Le Modèle Linéaire par l'exemple :

Régression, Analyse de la Variance et Plans d'Expériences

Illustrations numériques avec les logiciels R, SAS et Splus

Jean-Marc Azaïs et Jean-Marc Bardet

Introduction

Le modèle linéaire est souvent le premier outil de statistique inférentielle mis en œuvre. Il suit immédiatement l'étude descriptive des données. Son intérêt principal réside dans la simplicité de ses algorithmes d'estimation et de test qui permettent, sans problème majeur, de poser des modèles à plusieurs centaines de paramètres. Cette richesse lui donne une grande souplesse et, dans une certaine mesure, la capacité de s'adapter à la plupart des situations. Bien que, une fois le problème bien "débroussaillé", il doive parfois céder la place à des techniques plus sophistiquées, le modèle linéaire n'en reste pas moins une pierre fondamentale de l'édifice statistique.

L'enseignement et l'utilisation du modèle linéaire amènent à un paradoxe. En effet, les formules de base sont très peu nombreuses (voir le chapitre 3). Il semblerait à première vue qu'un document d'une vingtaine de pages soit suffisant. En fait, il n'en est rien, l'étude des problèmes concrets, la compréhension des sorties de logiciel, amènent à des questions dont les réponses découlent des formules fondamentales, mais de façon non-triviale. Il faut alors développer de nouvelles notions. Nous avons donc introduit les principes généraux qui permettent de s'adapter à chacune des situations, mais nous avons délibérément évité de prétendre donner un traitement explicite de toutes les situations les plus courantes liées au modèle linéaire. Une telle exhaustivité se payerait par une présentation très fastidieuse et très calculatoire.

Ce livre est donc une tentative de conciliation sur les principaux thèmes attenant au modèle linéaire, c'est-à-dire les notions de régression, d'analyse de la variance, de modèle mixte et de plans d'expériences, entre les propos très (voire trop) succincts que l'on peut trouver dans les livres classiques de statistique inférentielle (voir par exemple Dacunha-Castelle et Duflo [20] et [21], Milhaud [45], P.S. Toulouse [59], etc...) et les livres spécialisés sur une de ces notions (voir par exemple Guyon [31], Tomassone *et al.* [58], etc...). **De nombreux exemples sont présentés pour introduire et illustrer les résultats théoriques.** Ainsi, avant chaque propos un peu général ou abstrait, un ou plusieurs exemples simples permettent de se familiariser avec les notations, les questions posées, les traitements possibles, la problématique... **Un bon nombre de fins de chapitre comprennent une application des résultats sur**

des jeux de données traités par les logiciels statistiques SAS, Splus et R (pour plus de détails sur l'utilisation des logiciels, voir un peu après) et une séquence d'exercices dont les niveaux ont été subjectivement appréciés par les auteurs (de (*) à (***)). On trouvera des corrections de ces exercices ainsi que les bases de données proposées dans les applications informatiques sur le site : `http://www.dunod.com`. Un chapitre entier (chapitre 15) est dédié à trois études de cas concrètes traitées avec les logiciels, qui synthétisent et approfondissent une bonne part de ce qui a été vu dans le reste du livre. Tout ceci constitue un ouvrage où la théorie et les applications sont mutuellement au service l'un de l'autre (ce qui le distingue notamment du livre complet mais très abstrait de J. Coursol [19]).

Notons à propos que les applications informatiques ont pour but d'illustrer les notions développées en début de chapitre. En aucun cas nous ne prétendons donner une liste détaillée de toutes les possibilités des logiciels. Notre but serait plutôt de mettre en avant les incroyables possibilités qu'offrent ces logiciels, tout en éveillant la prudence de leurs utilisateurs. En effet, quoi qu'il advienne, un logiciel "sortira" des résultats numériques, des graphes. On pourrait alors facilement se reposer sur ces traitements rapides et puissants, en laissant finalement le logiciel réfléchir à sa place sans savoir véritablement les calculs et méthodes qu'il a mis en œuvre. Une telle attitude est dangereuse et conduit souvent à des résultats aberrants... Nous verrons même que les trois seuls logiciels que nous étudierons effectuent parfois des traitements différents avec des commandes semblables ! A nos yeux, il sera plus convaincant et efficace de n'utiliser qu'une faible partie des nombreuses commandes des logiciels, mais de bien les utiliser en connaissant leurs possibilités et limites.

Issu d'enseignements délivrés à des étudiants de niveau (licence, mastere, écoles d'ingénieur) et d'origines (mathématiques pures et appliquées, économétrie, biométrie) diverses, ce livre a été conçu pour permettre une lecture à plusieurs niveaux. A part quelques paragraphes parfois un peu plus denses, les chapitres 1 à 5 pour le modèle linéaire et les chapitres 11 à 13 pour la planification expérimentale, sont très accessibles et directement centrés sur les applications (avec également de nombreuses illustrations utilisant les logiciels statistiques). Ils se veulent accessibles à un lecteur peu féru de mathématiques. Dans le même ordre d'idée, un rappel des notions de bases minimales de la théorie des probabilités est donné en annexe. Les autres chapitres peuvent contenir des démonstrations ayant un certain contenu mathématique. En particulier les chapitres 9 et 12 sont issus de cours de mastere recherche en mathématiques appliquées. Cependant, les thèmes abordés dans ces mêmes chapitres nous paraissent fondamentaux pour les applications : l'abstraction des concepts et démonstrations proposés n'est donc pas gratuite.

Ainsi nous pensons que **des étudiants provenant de formations aussi variées**

que des licences (3ème année) ou masteres spécialité recherche ou professionnelle, en mathématiques, économie, chimie, biométrie, biostatistique, etc... ou en école de chimie, d'agronomie ou de commerce, pourront trouver profit à la lecture de cet ouvrage. Par ailleurs, **ce livre est également destiné aux statisticiens d'entreprise, qui pourront en particulier se reporter avantageusement aux nombreuses applications et illustrations informatiques.**

On pourrait grossièrement décomposer ce texte en deux parties de tailles assez inégales. La première partie est directement centrée sur le modèle linéaire et l'analyse de la variance, puisque, nous le répéterons de nombreuses fois, celle-ci est un cas particulier du modèle linéaire. Seront ainsi abordés les principaux résultats théoriques à retenir en ce qui concerne le modèle linéaire gaussien et (surtout) non gaussien, mais aussi l'utilisation de ses résultats, leurs limitations et extensions possibles. Il se révélera alors nécessaire d'évoquer les problèmes de sélection de modèle en régression linéaire, que ce soit dans un cadre explicatif ou prédictif. La seconde partie donne un exposé succinct de la théorie des plans d'expériences : comment optimiser la qualité des données en vue de leur utilisation par un modèle linéaire. Cette partie est fondamentale pour les applications. Les gains que l'on peut y réaliser sont parfois spectaculaires et souvent supérieurs au gain apporté par l'utilisation d'une méthode sophistiquée en lieu et place d'une méthode dite standard. Un chapitre plus théorique (chapitre 12) donne une présentation "à l'anglaise" des décompositions d'expériences en strates. Cette partie relativement technique nous a paru indispensable car elle montre de manière rigoureuse le lien entre la randomisation et le modèle d'analyse.

Pour terminer, évoquons ce que l'on ne trouvera pas dans ce livre pour des raisons de concision : des extensions assez naturelles comme la régression logistique et le modèle linéaire généralisé ; la régression fonctionnelle (ou non-paramétrique) ; des sujets pouvant avantageusement précéder ou compléter l'exploitation d'un modèle linéaire et regroupés dans ce que l'on appelle désormais le data mining (exploration des données), par exemple l'analyse en composantes principales, l'analyse factorielle, la classification,... ; enfin des extensions plus lointaines mais parfois essentielles pour améliorer les résultats obtenus avec un modèle linéaire, comme les modèles non-linéaires, les réseaux de neurones, les modèles CART (arbres de régression), les méthodes de bootstrap, de boosting, de bagging, etc... Pour beaucoup de ces thèmes "oubliés", on pourra se reporter aux documents en ligne de de P. Besse [13], [11] et [12].

Nous vous souhaitons une très bonne lecture.

Toulouse-Paris Septembre 2005

Jean-Marc Azaïs et Jean-Marc Bardet

Remerciements

Ce livre n'aurait pas été possible sans l'aide de nombreuses personnes. Il est sans-doute difficile de toutes les citer. Nous tenons à remercier la formation permanente de l'INRA qui a initié ce projet. Beaucoup d'échanges d'information, de polycopiés, d'exemples ont eu lieu avec nos collègues de l'INRA et de l'Université Toulouse III, en particulier Alain Baccini, Bernard Bercu, Philippe Besse, Christine Durier, Jean-Claude Fort, Anne-Laure Fougères, Fabrice Gamboa, Xavier Guyon (pour l'université Paris I), André Kobilinski, Béatrice Laurent, Hervé Monod, Clémentine Prieur et Henri Caussinus. Ce dernier nous a mis en contact avec Stephen Stiegler qui par sa connaissance de l'histoire de la statistique nous a aidé dans la rédaction du chapitre 2. Enfin ce livre doit beaucoup à nos étudiants qui nous ont donné l'envie et l'énergie de réaliser un tel ouvrage.

Notations

Le lecteur se reportera à l'appendice pour tous les rappels concernant la théorie des probabilités et des statistiques. Nous donnons ici les seules notations qui sont indispensables à la compréhension de l'ouvrage. On se placera en général dans la base canonique de $\mathbb{R}^d$ muni du produit scalaire euclidien standard $< ., . >$. Ainsi,

- de manière générale X et M correspondront plutôt à une matrice, les vecteurs seront notés par des majuscules romaines de la fin de l'alphabet (par exemple, Y, Z,...) ou par des lettres grecques (par exemple, θ, γ,...) sans indice. Les scalaires sont notés plutôt par des minuscules latines (par exemple, x ou a) ou bien par des majuscules romaines ou des lettres grecques indicées (par exemple, X_i ou θ_j), mais le lecteur devra parfois faire appel au contexte.

- si $Z_1, \cdots, Z_n$ sont n vecteurs à valeurs dans $\mathbb{R}^d$, $[Z_1, \cdots, Z_n]$ désignera le sous-espace vectoriel de $\mathbb{R}^d$ engendré par $Z_1, \cdots, Z_n$.

- la matrice M à n lignes et p colonnes dont l'élément ij vaut M_{ij} sera notéé :

$$M = \{M_{ij}, i = 1, ..., n, j = 1, ..., p\}$$

- si M est une matrice, nous noterons $[M]$ l'espace vectoriel engendré par les vecteurs représentés par les colonnes de M.

- si E est un sous-espace vectoriel de $\mathbb{R}^d$, nous noterons P_E le projecteur orthogonal sur E pour le produit scalaire $< ., . >$, c'est-à-dire que $P_E(Z)$ est le seul vecteur de $\mathbb{R}^d$ qui vérifie

$$P_E(Z) \in E \quad \text{et pour tout} \quad Y \in E, \ < Z, Y >=< P_E(Z), Y > .$$

  Cette notation $P_E(Z)$ désignera aussi bien le projecteur (endomorphisme de $\mathbb{R}^d$) que la matrice associée (dans la base canonique de $\mathbb{R}^d$).

- X' est la matrice transposée de la matrice X.

- Id est la matrice carrée identité et $\mathbb{1}$ le vecteur dont toutes les coordonnées sont égales à 1.

Principes généraux d'utilisation des logiciels

Dans la plupart des chapitres qui suivent, nous allons faire appel à des illustrations informatiques à l'aide de trois logiciels, SAS , Splus et R . Ce choix de logiciels s'explique par des critères de popularité et de performance. Le logiciel R, peut-être un peu moins convivial que les deux autres, a la particularité d'être gratuit (c'est un clone du langage S sous licence GNU) et téléchargeable à partir des sites `http://www.r-project.org/` ou `http://cran.cict.fr/`. De plus, ce logiciel est enrichi par des "packages" que développent en permanence et gracieusement des chercheurs du monde entier.

Pour chaque traitement, nous avons choisi de retranscrire les commandes détaillées propres à chaque logiciel. Il est clair que, tout au moins sur les versions les plus récentes de ces logiciels, une bonne part des traitements auraient pu être directement obtenus en quelques "clics" de souris (on peut penser notamment à la convivialité apportée par SAS/INSIGHT ou Splus 2000). Cependant, nous avons préféré les commandes tapées "à l'ancienne", commandes pour lesquelles toutes les options doivent être précisées "à la main", et cela pour quatre raisons : d'abord, parce que cela donne tout de même un peu plus de contrôle sur ce que l'on fait, ensuite, parce que certaines possibilités n'existent qu'avec des commandes tapées (on pense par exemple à la sélection de variables), parce que cela permet de travailler avec d'anciennes versions des logiciels ou avec des systèmes d'exploitation moins conviviaux et enfin parce que cela pourrait permettre d'automatiser ou de réexécuter rapidement des commandes complexes sur différents échantillons en ne modifiant que quelques paramètres. On trouvera de nombreux documents pédagogiques sur SAS, Splus ou R, sur le site `http://www.lsp.ups-tlse.fr/`.

Voyons maintenant en détail quelques principes généraux relatifs aux trois logiciels, puis ceux spécifiques à chacun d'eux.

Quelques propos concernant les trois logiciels utilisés

- Comme il a été précisé un peu plus haut, les commandes devront être "écrites" dans la fenêtre de commandes (appelée `Commande Window`, `Console` ou `Editor`) suivant les logiciels, leurs versions et les systèmes d'exploitation (Linux, Unix, Mac, Windows,...).

- Toutes les noms de commandes, de procédures, de fichiers, de variables..., rela-

tifs à un logiciel seront écrits dans les chapitres qui suivent avec la typographie suivante : `lm`, `proc reg`, `foret`, `temperature`,... (comme les instructions informatiques n'admettent pas les accents, il ne faudra pas s'offusquer de voir apparaître quelques fautes d'orthographe en ce qui concerne ces noms...).

- De nombreuses illustrations feront appel à des fichiers de données que l'on pourra pour la plupart télécharger (à partir du site `http://www.dunod.com`).

- La démarche suivie dans les chapitres qui suivent est de retranscrire stricto sensu les instructions à "écrire" sur la fenêtre de commande (il peut être intéressant de les écrire a priori dans un fichier texte et de travailler avec des `Copier/Coller`...). Cela permet également un apprentissage progressif du (des) logiciel(s). Cependant, il est souhaitable d'avoir déjà quelques notions minimales. Par ailleurs, du fait d'une certaine progression des chapitres, il est souhaitable de commencer par les illustrations des premiers chapitres avant de traiter celles des suivants.

- Le chapitre 15 sera uniquement consacrée à des études de cas (issues de données réelles) reprenant ou prolongeant l'ensemble des chapitres. Un seul logiciel sera utilisé par étude et les instructions ne seront pas toutes retranscrites.

Principes généraux relatifs au logiciel SAS

Le traitement d'un exemple avec le logiciel SAS comprend en général deux étapes (avec ce logiciel plusieurs commandes de suite peuvent être écrites avant d'être "soumises", c'est-à-dire traitées par le logiciel en même temps). Bien que rien n'empêche de soumettre donc de réaliser ces deux étapes en même temps, il est conseillé de le faire séparément, ce qui cela permet de mieux détecter les erreurs que vous ne manquerez pas de faire au départ :

- une étape "data" qui consiste à insérer dans SAS un jeu de données, soit en "rentrant directement à la main" ces données par la procédure `proc data`, soit en important un fichier texte (ou `Access`, `Excel`,...), et à nommer ou renommer les variables (si elles ne l'ont pas déjà été). Le résultat de cette étape est la création d'un tableau de données SAS. Notons également que toute variable numérique est considérée par SAS a priori comme quantitative, sauf si elle suit une commande `class`, auquel cas elle devient qualitative.

- une étape "procédure" qui prend les données d'un tableau de données SAS (instruction `data=sasuser.foret3` par exemple) et qui effectue l'analyse statistique proprement dite. Les principales procédures en rapport avec ce document sont

(il en existe bien d'autres concernant le modèle linéaire, comme `proc anova`, `proc probit`,..., mais pour des raisons pédagogiques nous avons préféré nous restreindre à celles-ci) :

- `proc reg` pour la régression ;
- `proc glm` pour les modèles linéaires en général ;
- `proc mixed` pour les modèles mixtes ;
- `proc plan` pour la génération de plans d'expériences ;
- `proc factex` pour la génération et l'analyse de plans d'expériences factoriels ;
- `proc optex` pour la recherche de plans optimaux.

Principes généraux relatifs aux logiciels R et Splus

Le traitement d'un exemple en R ou Splus (comme nous l'avons déjà évoqué, les deux langages sont très proches) comprend également deux parties (les commandes peuvent être écrites et traitées par le logiciel les unes après les autres ou bien regroupées dans un fichier) :

– une étape "data" qui consiste à insérer dans le logiciel (et plus particulièrement dans le `workspace`, espace de travail, que l'on peut sauvegarder) un jeu de données, soit en "rentrant directement à la main" ces données, soit en important un fichier texte (ou `Access`, `Excel`,...), et à nommer les variables (si elles ne l'ont pas déjà été). Plusieurs types d'objets peuvent ainsi avoir été crées :

  - un objet `vector` qui est un vecteur de données numériques, auquel on donne un nom (par exemple `X`) ;
  - un objet `matrix` qui est une matrice de données numériques, à laquelle on donne un nom (par exemple `M`) et dont les colonnes ou les lignes peuvent également avoir un nom ;
  - un objet `data.frame` qui est un tableau de données (numériques ou qualitatives), auquel on donne un nom (par exemple `beton`) ;
  - un objet `list` qui est une liste de différents autres objets. Un `data.frame` est un objet `list` particulier.

  D'autres types d'objets existent également (par exemple `ts`, pour "times series") mais ne seront pas utilisés ici. Pour connaître le type d'objet que l'on manipule, on peut "taper" la commande `is.vector(X)` ou `is.data.frame(proc)`. Pour passer, lorsque cela est possible, d'un type d'objet à un autre, on peut taper par exemple `M.frame<-as.data.frame(M)`. Enfin, il sera souvent bienvenu de vérifier que ce que l'on voulait être numérique l'est réellement : une réalisation

d'une variable qualitative (ou alphanumérique) apparaît entre des guillemets (par exemple, `"vert"`). Pour savoir si l'on a à faire à des données quantitatives ou qualitatives, on pourra taper la commande `class` suivi du nom de la variable. Enfin, pour que des données numériques puissent être considérer comme des réalisations de variables qualitatives, on tapera la commande `as.factor`.

– une étape de traitement des données faisant appel à la base de données précédemment créée (le plus souvent ce sera un `data.frame`). Les commandes de traitement en rapport avec ce document sont :

- `lm` (R et Splus) ou `menuLm` (Splus) pour la régression ;
- `glm` pour les modèles linéaires généralisés ;
- `Anova` pour l'analyse de variance et de covariance (à préférer en général à la commande `anova`) en R ;
- `leaps`, `AIC`, `BIC`, `stepAIC` pour la sélection de modèles ;
- `sample` pour la construction de plans d'expériences.
- `menuFacDesign` (Splus) ou `gen.factorial` (R) pour la construction de plans fractionnaires.
- `optFederov` pour la recherche de plans optimaux (en R).

Chapitre 1

Exemples Simples

Dans ce chapitre, nous rappelons brièvement les formules de la régression linéaire simple et de l'analyse de la variance à un facteur. Notre but est de faire ressortir la similitude des hypothèses et des méthodes pour faire apparaître la nécessité de les traiter en détail dans un même cadre. C'est ce qui sera fait au chapitre 3. En attendant cette étude globale, nous allons donner de nombreux résultats sans justification.

1 Régression linéaire simple

1.1 Exemple

On considère 5 groupes de femmes âgées respectivement de 35, 45, 55, 65 et 75 ans. Dans chaque groupe, on a mesuré la tension artérielle en mm de mercure de chaque femme et on a calculé la valeur moyenne pour chaque groupe. On définit donc les variables :

Y : tension moyenne en mm Hg	114	124	143	158	166
Z : âge du groupe considéré	35	45	55	65	75

(source de ces données : Snedecor et Cochran [55] p. 136)

Afin de visualiser ces données, on fait une représentation cartésienne (nous verrons un peu plus loin comment obtenir simplement une telle figure) :

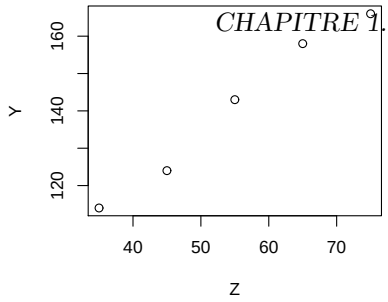


FIGURE 1.1 – Nuage de points des couples (Z_i, Y_i)

Commentaires : Sur le graphique on constate que

- la tension artérielle augmente avec l'âge (résultat bien classique).
- mais surtout, cette augmentation semble linéaire puisque les points du graphique sont presque alignés.

1.2 Modèle et estimation

Notons Y_i la tension artérielle du $i$ème groupe et Z_i son âge, nous pouvons alors proposer le modèle suivant :

$$Y_i = \mu + \beta \cdot Z_i + \varepsilon_i. \tag{1.1}$$

C'est un modèle de dépendance linéaire. Il y a deux paramètres, μ est appelé la *constante* (*"Intercept"* en anglais), β la *pente*. Ils sont tous les deux inconnus. Le vecteur aléatoire ε formé par les variables aléatoires ε_i est appelé *l'erreur du modèle*.

Plus généralement, supposons que l'on a n observations connues (dans l'exemple, $n = 5$) d'une variable Y appelée variable à expliquer (dans l'exemple, la tension artérielle) et d'une variable Z dite explicative (dans l'exemple, l'âge). On supposera de plus que pour $i = 1, \cdots, n$, les variables Y_i et Z_i suivent le modèle (1.1) et que

- pour $i = 1, \cdots, n$, $\mathbb{E}(\varepsilon_i) = 0$ (les erreurs sont centrées) ;
- pour $i = 1, \cdots, n$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (donc la variance des erreurs est constante) et nous supposerons que σ^2 est un autre paramètre inconnu ;
- pour $i = 1, \cdots, n$, les variables ε_i sont indépendantes et de loi gaussienne (dite encore loi normale).

Ces postulats seront commentés un plus loin.

Remarque : Par abus de notation, on désignera aussi bien par Y la variable statistique que le vecteur $(Y_i)_{1\leq i\leq n}$. Le contexte permettra en général de distinguer entre les deux cas.

Pour déterminer les paramètres inconnus μ et β (ainsi également que σ^2), une méthode possible est la méthode dite des moindres carrés (ordinaires). Celle-ci consiste d'abord à déterminer des valeurs m et b minimisant la fonction :

$$SCR(m,b) := \sum_{i=1}^{n}[Y_i - (m + bZ_i)]^2.$$

Cela revient à minimiser les carrés des écarts pris verticalement entre la droite de paramètres m et b (c'est-à-dire la droite $y = m+b\cdot z$) et les différents points observés. La solution bien connue de ce problème (on peut la retrouver en dérivant $SCR(m,b)$ par rapport à m et à b) fournit les valeurs $(\widehat{\mu},\widehat{\beta})$, dites solutions des *moindres carrés ordinaires*, avec :

$$\begin{cases} \widehat{\beta} &= \dfrac{\sum_{i=1}^{n}(Y_i-\overline{Y})(Z_i-\overline{Z})}{\sum_{i=1}^{n}(Z_i-\overline{Z})^2} \\ \\ \widehat{\mu} &= \overline{Y} - \widehat{\beta}\cdot\overline{Z} \end{cases}$$

où $\overline{Z} := \dfrac{1}{n}\sum_{i=1}^{n} Z_i$ est la moyenne des Z_i et $\overline{Y} := \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ la moyenne des Y_i. On définit également :

- le vecteur des valeurs estimées $\widehat{Y} = \widehat{\mu} + \widehat{\beta}\cdot Z = (\widehat{\mu} + \widehat{\beta}\cdot Z_i)_{1\leq i\leq n}$;
- le vecteur des *résidus* $\widehat{\varepsilon} = Y - \widehat{Y} = (Y_i - (\widehat{\mu} + \widehat{\beta}\cdot Z_i))_{1\leq i\leq n}$;
- l'estimateur de la variance $s^2 = \widehat{\sigma^2} = \dfrac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$. Le coefficient $n-2$ peut s'expliquer par la règle : nombre de données (ici n) moins le nombre de paramètres du modèle (ici 2). Nous verrons un peu plus loin la justification d'une telle renormalisation.

Remarque 1 : Par la suite et pour aider à la lecture des résultats, si θ est un vecteur de paramètres réels inconnus (par exemple σ^2 ou (μ,β)), nous adopterons la convention de noter $\widehat{\theta}$ un estimateur de θ.

Remarque 2 : Pourquoi utiliser cette sommes des moindres carrés, c'est-à-dire la

somme des distances quadratiques suivant l'axe des ordonnées ? On utilise parfois d'autres fonctions comme par exemple :

- la somme des valeurs absolues en posant

$$SVA(m,b) := \sum_{i=1}^{n} |Y_i - (m + b \cdot Z_i)|. \tag{1.2}$$

  La régression aux moindres valeurs absolues (minimisation de la formule (1.2)) peut s'utiliser quand on suspecte la présence de quelques *données aberrantes* (appelées encore valeurs atypiques, *outliers* en anglais, et qui correspond soit à une erreur dans la mesure ou la retranscription de la donnée, soit à un individu possédant des caractéristiques très singulières par rapport au reste de la population), mais elle est complexe à mettre en place numériquement.

- la somme des distances euclidiennes (classiques) entre les points et la droite (cf. figure 1.2), soit :

$$SDE(m,b) := \sum_{i=1}^{n} d(M_i, \Delta_{m,b}), \tag{1.3}$$

  où les M_i sont les points de $\mathbb{R}^2$ de coordonnées (Z_i, Y_i), $\Delta_{m,b}$ est la droite d'équation cartésienne $y = m + b \cdot z$ et $d(M_i, \Delta_{m,b})$ représente la distance euclidienne entre M_i et la droite $\Delta_{m,b}$, soit encore $d(M_i, \Delta_{m,b}) = d(M_i, P_{\Delta_{m,b}}(M_i))$, $P_{\Delta_{m,b}}(M_i)$ étant le projeté orthogonale de M_i sur la droite $\Delta_{m,b}$. La solution d'un tel problème, et sa généralisation à la régression linéaire multiple, fait appel à des techniques de diagonalisation de la matrice de covariance empirique et d'ordonnancement des valeurs propres de cette matrice. Une telle méthode se rapproche donc de l'analyse en composantes principales (voir par exemple Saporta [52]). On l'utilisera quand les deux variables, Y et Z, jouent des rôles symétriques, par exemple la longueur et la largeur d'une feuille de plante. Ici dans l'exemple des tension artérielles, ce n'est pas le cas. La variable `age` est parfaitement déterminée et c'est bien elle qui influe sur la variable `tension`. Notons au passage que pour cette raison nous avons choisi l'exemple des tension artérielles plutôt que l'exemple historique de Karl Pearson sur les tailles des pères Z et de leurs fils Y dont la dissymétrie est moins évidente. Ce dernier exemple a donné lieu à la première utilisation du mot "régression" : dans l'analyse, le coefficient $\widehat{\beta}$ est inférieur à 1, c'est-à-dire que si deux pères ont 10 cm d'écart, l'écart de tailles entre leurs fils est moindre : il y a *régression* des différences. Sur 1078 familles retenues, Pearson a trouvé numériquement la valeur $\beta \simeq 0.516$. On trouve cet exemple dans le livre de Snedecor et Cochran [55]

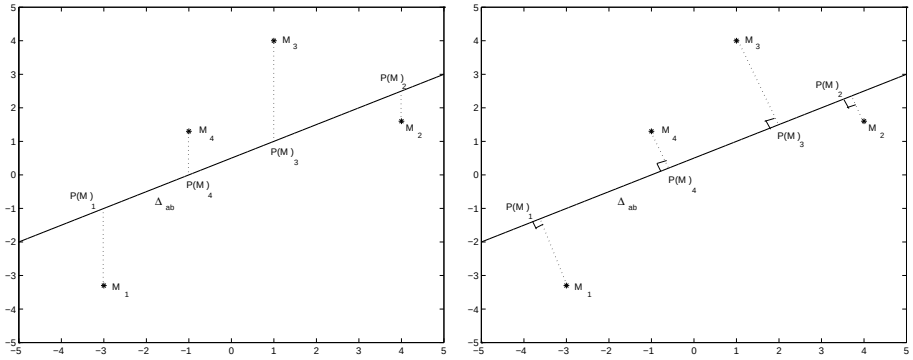


FIGURE 1.2 – Distances des points (Z_i, Y_i) à la droite Δ_{mb} suivant : 1/ une projection parallèlement à l'axe des ordonnées ($\longrightarrow$ régression linéaire classique par moindres carrés) 2/ une projection orthogonale ($\longrightarrow$ régression par minimisation de la sommes des carrés des distances euclidiennes)

L'approche "régression linéaire", avec estimation par moindres carrés, est devenu classique essentiellement pour deux raisons : (1) les solutions sont explicites et de faible complexité numérique, même pour des modèles beaucoup plus complexes que le modèle 1.1 ; (2) ce choix est optimal pour des observations gaussiennes (comme on le verra au chapitre 3).

En se plaçant dans le cadre du modèle (1.1) et en considérant les X_i comme des données déterministes (connues), les Y_i sont des variables aléatoires gaussiennes. On peut alors apprécier la précision des estimateurs $\widehat{\mu}$ et $\widehat{\beta}$ à l'aide des formules complémentaires suivantes :

- $\mathbb{E}(\widehat{\mu}) = \mu$ et $\mathbb{E}(\widehat{\beta}) = \beta$;
- $\mathrm{Var}(\widehat{\beta}) = \dfrac{\sigma^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$ et $\mathrm{Var}(\widehat{\mu}) = \sigma^2 \left(\dfrac{1}{n} + \dfrac{\overline{Z}^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2} \right) = \dfrac{\sigma^2}{n} \dfrac{\sum_{i=1}^n Z_i^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$;
- $\mathrm{cov}(\widehat{\alpha}, \widehat{\beta}) = -\dfrac{\sigma^2 \cdot \overline{Z}}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$.

De la même manière, on définit la matrice de variance-covariance du vecteur $(\widehat{\mu}, \widehat{\beta})$ qui vérifie :

$$\mathrm{Var}\begin{pmatrix} \widehat{\mu} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(\widehat{\mu}) & \mathrm{cov}(\widehat{\mu}, \widehat{\beta}) \\ \mathrm{cov}(\widehat{\mu}, \widehat{\beta}) & \mathrm{Var}(\widehat{\beta}) \end{pmatrix} = \frac{\sigma^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n Z_i^2 & -\overline{Z} \\ -\overline{Z} & 1 \end{pmatrix}.$$

1.3 Table d'analyse de la variance

On complète l'étude précédente en construisant la table suivante, encore appelée *table d'analyse de la variance* :

Source	Somme de carrés	Degrés de liberté	Carré moyen	$\widehat{F}$
régression	$\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$	1	$\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$	
résiduelle	$\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$	$n-2$	$\frac{1}{n-2} \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$	$(n-2)\ \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}$
totale	$\sum_{i=1}^n (Y_i - \overline{Y})^2$	$n-1$	$\frac{1}{n-1} \sum_{i=1}^n (Y_i - \overline{Y})^2$	

Commentaires :

- la statistique $\widehat{F}$, dite *statistique de Fisher* qui permet de tester la nullité de la pente, à savoir $\beta = 0$, est égale au rapport entre le carré moyen expliqué part la régression et le carré moyen résiduel. Plus précisément, cela ce traduit dans notre modèle par le fait que l'on va tester l'hypothèse :

$$\text{contre} \quad \begin{array}{rcl} H_0 & : & \beta = 0 \\ H_1 & : & \beta \neq 0 \end{array} .$$

  L'égalité $\beta = 0$ permet de définir un *sous-modèle* du modèle (1.1). Pour un test de niveau α (en général $\alpha = 5\%$), on compare la statistique $\widehat{F}$ à la valeur dépassée avec une probabilité α par une variable aléatoire distribuée suivant la loi de Fisher à $(1, n-2)$ degrés de liberté. Cette quantité, notée $F_{(1,n-2),1-\alpha}$ est le quantile d'ordre $(1-\alpha)$ de cette loi de Fisher à $(1, n-2)$ degrés de liberté.

- la somme des carrés résiduelle est le minimum de $SCR(m, b)$, soit

$$SCR(\widehat{\mu}, \widehat{\beta}) = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2.$$

- la somme des carrés expliquée par la régression, $\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2$, est la quantité expliquée par la droite de régression par rapport au modèle où on n'ajuste les données qu'avec une simple moyenne $\overline{Y}$ (ce qui revient à faire une régression sur une droite de pente nulle).

– la somme des carrés totale est normalement utilisée pour le calcul de la variance empirique.

Remarque : Pour mesurer l'adéquation d'un modèle linéaire aux données, le *coefficient de détermination*, ou coefficient R^2, est souvent proposé. Sa définition est la suivante :

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_i)^2} = \frac{\sum_{i=1}^n (\overline{Y}_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_i)^2}. \tag{1.4}$$

Intuitivement, on comprend bien que lorsque la régression est "précise", alors la variance empirique des résidus est négligeable devant la variance empirique des réponses, donc le coefficient R^2 est proche de 1. On verra au chapitre 3 une utilisation de ce coefficient de détermination dans des modèles linéaires très généraux. Ici, dans le cas d'un seul régresseur, ce coefficient R^2 peut également s'écrire comme le carré du coefficient de corrélation (empirique), soit $R^2 = \rho^2 = \left(\frac{\overline{\sigma}_{ZY}}{\overline{\sigma}_Y \cdot \overline{\sigma}_Z} \right)^2$, où $\overline{\sigma}_{YZ}$ est la covariance empirique entre Y et Z, et $\overline{\sigma}_Y$, $\overline{\sigma}_Z$ sont les écart-types empiriques de Y et de Z.
Cependant, même si, dans ce cas précis de la régression linéaire par rapport à une variable, on peut obtenir une loi asymptotique de ce coefficient R^2 (voir par exemple Fisher [28]), on préférera utiliser les statistiques de Fisher plutôt que le coefficient R^2 pour tester l'adéquation de la régression linéaire (voir également la remarque du chapitre suivant 3).

1.4 Intervalle de confiance et Test de Student

D'après ce qui précède $\widehat{\beta}$ est une variable gaussienne centrée et comme $\mathrm{Var}\,(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$, un estimateur naturel de cette variance est $\frac{\widehat{\sigma^2}}{\sum_{i=1}^n (Z_i - \overline{Z})^2}$. Après une renormalisation par la racine de cette variance empirique, on montre alors que

$$\widehat{T} = \frac{\widehat{\beta}}{\sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^n (Z_i - \overline{Z})^2}}} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(Z_i - \overline{Z})}{\sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 \sum_{i=1}^n (Z_i - \overline{Z})^2}} \tag{1.5}$$

est une variable aléatoire distribuée selon une loi de Student à $(n-2)$ degrés de liberté (loi notée T_{n-2} et dite *statistique de Student*). On déduit de ceci qu'un intervalle de confiance de niveau $(1-\alpha)$ sur la valeur de β est

$$\left[\widehat{\beta} - T_{n-2,1-\alpha/2} \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}} \ , \ \widehat{\beta} + T_{n-2,1-\alpha/2} \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^n (Z_i - \overline{Z})^2}} \right]$$

où $T_{n-2,1-\alpha/2}$ est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $(n-2)$ degrés de liberté (voir l'annexe pour une définition) défini par

$$\Pr(X \leq T_{n-2,1-\alpha/2}) = 1-\alpha/2,$$

pour X de loi T_{n-2}.

Il est également possible de réaliser un *test de Student* de niveau α sur la nullité de la pente. Plus précisément, on testera à l'aide de la statistique $\widehat{T}$, l'hypothèse

$$\text{contre} \quad \begin{array}{rcl} H_0 & : & \beta = 0 \\ H_1 & : & \beta \neq 0 \end{array}.$$

Ce test revient à regarder si la valeur zéro appartient à l'intervalle de confiance ci-dessus. En conséquence, si

$$|\widehat{\beta}| \sqrt{\frac{\sum_{i=1}^n (Z_i - \overline{Z})^2}{\widehat{\sigma^2}}} > T_{n-2,1-\alpha/2},$$

l'hypothèse H_0 sera rejeté (on dit alors que le test est significatif), sinon l'hypothèse H_0 sera acceptée (on dit alors que le test n'est pas significatif). Or pour p quelconque, le carré d'une variable de Student à p degrés de liberté est une variable de Fisher à $(1,p)$ degrés de liberté. Par suite, le test présenté un peu plus haut et mettant en œuvre la statistique $\widehat{F}$, est le même que celui portant sur la statistique $\widehat{T}$. Dans le cadre ici présenté (et cela sera généralisé un peu plus loin),

le test de Student est donc strictement le même que le test de Fisher issu de la table d'analyse de la variance.

2 Analyse de la variance à un facteur

2.1 Exemple

Un forestier s'intéresse aux hauteurs moyennes de trois forêts. Pour les estimer, il échantillonne un certain nombre d'arbres et mesure leurs hauteurs :

forêt1	$n_1 = 6$	23,4 24,4 24,6 24,9 25,0 26,2
forêt2	$n_2 = 5$	22,5 22,9 23,7 24,0 24,0
forêt3	$n_3 = 7$	18,9 21,1 21,1 22,1 22,5 23,5 24,5

(source de ces données : Dacunha-Castelle et Duflo [20])

Ces données peuvent être présentées de deux manières équivalentes :

i. On dispose de trois échantillons indépendants et on désire comparer leurs moyennes. C'est la présentation élémentaire dite de "comparaison de moyennes".

ii. On dispose d'un seul échantillon de longueur 18 et d'une variable explicative qualitative, ou facteur, le numéro de la forêt. En prenant ce second point de vue, on parle d'analyse de la variance à 1 facteur. C'est également le point de vue adopté par la plupart des logiciels et il offre l'avantage de s'adapter à des cas compliqués : 2 facteurs, 3 facteurs, etc... Et c'est également le point de vue que nous adopterons le plus souvent.

2.2 Modèle statistique

La méthode de collecte des données, à savoir un échantillonnage indépendant dans chacune des forêts, nous permet de proposer le modèle suivant :

- on note Y_{ij} la hauteur du $j^{\text{ème}}$ arbre de la forêt i ;
- on note μ_i la hauteur moyenne de la forêt i (que l'on pourrait théoriquement calculer en recensant tous les arbres qui s'y trouvent).

Dans ce cadre, un modèle possible est le suivant :

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{1.6}$$

où ε_{ij} est la variabilité de l'arbre j par rapport à la hauteur moyenne de la forêt i. Comme précédemment, nous allons faire quelques hypothèses sur les ε_{ij} :

- on a par définition $\mathbb{E}(\varepsilon_{ij}) = 0$, pour tout (i, j) ;
- on suppose que la variance de la hauteur des arbres est la même dans chaque forêt, soit $\text{Var}(\varepsilon_{ij}) = \sigma^2$, pour tout (i, j) ;
- l'échantillonnage tel qu'il a été effectué implique que les mesures et donc les ε_{ij} sont indépendants ;
- enfin, si les effectifs sont petits (ce qui est le cas dans cet exemple forestier), on supposera que les Y_{ij} (et donc les ε_{ij}) sont des variables gaussiennes.

La question cruciale (mais ce n'est pas forcément la seule) à laquelle nous voudrions répondre est :

"les forêts sont-elles équivalentes ?"

Cela ce traduit dans notre modèle par le fait que l'on va tester l'hypothèse :

$$\text{contre} \quad \begin{array}{ll} H_0 : & \text{le modèle vérifie } \mu_1 = \mu_2 = \mu_3 \\ H_1 : & \text{le modèle ne vérifie pas } \mu_1 = \mu_2 = \mu_3 \end{array} .$$

L'égalité $\mu_1 = \mu_2 = \mu_3$ permet de définir un *sous-modèle* du modèle (1.6). En notant μ la valeur commune de μ_1, μ_2 et μ_3, ce sous-modèle s'écrit

$$Y_{ij} = \mu + \varepsilon_{ij}, \tag{1.7}$$

pour $i = 1, \cdots, I$ et $j = 1, \cdots, n_i$. Dans le grand modèle (1.6), on estimera la hauteur de la forêt i par la moyenne empirique de l'échantillon. Ainsi, si n_i est le nombre d'arbres de la forêt i, alors pour tout i :

$$Y_{i.} = \widehat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

Notation : dans tout ce qui suit, nous adopterons la notation suivante : un point à la place d'un indice veut dire la moyenne sur l'indice considéré.

• **Dans le "grand" modèle (1.6)**, on déduit de ce qui précède que,

- pour chaque (i, j), la valeur prédite de Y_{ij} est $\widehat{Y}_{ij} = Y_{i.} = \widehat{\mu}_i = \dfrac{1}{n_i} \displaystyle\sum_{j=1}^{n_i} Y_{ij}$;
- les résidus (estimation des erreurs ε_{ij}) sont les $\widehat{\varepsilon}_{ij} = Y_{ij} - Y_{i.}$
- la somme des carrés résiduelle (c'est-à-dire non prédite par le grand modèle (1.6), encore appelée somme des carrés totale) est donc

$$\mathrm{SC}_1 = \sum_{i,j} \widehat{\varepsilon}_{ij}^2 = \sum_{i,j} (Y_{ij} - Y_{i.})^2.$$

• **Dans le "sous-modèle" (1.7)** correspondant à l'hypothèse d'équivalence des forêts, on estime la moyenne commune μ par la moyenne empirique, soit :

$$\widehat{\mu} = \overline{Y} = Y_{..} = \frac{1}{n_1 + n_2 + n_3} \sum_{i,j} Y_{ij}.$$

Remarquons que dans le cas d'effectifs inégaux, cette quantité n'est pas égale à la moyenne des $Y_{i.}$.

La somme des carrés non prédite par le sous-modèle (1.7) (somme des carrés totale), est donnée par :

$$\mathrm{SC}_0 = \sum_{i,j} (Y_{ij} - \overline{Y})^2$$

• **Table d'analyse de la variance.** Par une utilisation de l'identité $(a+b)^2 = a^2 + 2ab + b^2$ ou par le théorème de Huygens, on obtient :

$$\mathrm{SC}_0 - \mathrm{SC}_1 = \sum_{i,j} (Y_{i.} - \overline{Y})^2 = \sum_{i,j} (\widehat{Y}_{ij} - \overline{Y})^2$$

(cette dernière écriture rappelant les résultats obtenus avec l'exemple de régression linéaire simple). En se plaçant dans le cadre général de I groupes, d'un effectif n_i pour le groupe i, et d'un nombre total de données $n = n_1 + \cdots + n_I$, on construit la table d'analyse de la variance :

Source	Somme de carrés	Degrés de liberté	Carré moyen	$\widehat{F}$
modèle	$\mathrm{SC}_0 - \mathrm{SC}_1 = \sum_{i,j}(Y_{i.} - Y_{..})^2$	$I-1$	$\frac{1}{I-1}\sum_{i,j}(Y_{i.} - Y_{..})^2$	
résiduelle	$\mathrm{SC}_1 = \sum_{i,j}(Y_{ij} - Y_{i.})^2$	$n-I$	$\frac{1}{n-I}\sum_{i,j}(Y_{ij} - Y_{i.})^2$	$\frac{n-I}{I-1}\frac{\sum_{i,j}(Y_{i.} - Y_{..})^2}{\sum_{i,j}(Y_{ij} - Y_{i.})^2}$
totale	$\mathrm{SC}_0 = \sum_{i,j}(Y_{ij} - \overline{Y})^2$	$n-1$	$\frac{1}{n-1}\sum_{i,j}(Y_{ij} - \overline{Y})^2$	

Remarque : certaines des doubles sommes présentées pourraient être écrites comme des sommes simples. Ainsi, $\sum_{i,j}(Y_{i.} - Y_{..})^2 = \sum_i n_i \cdot (Y_{i.} - Y_{..})^2$. Cependant, on préférera l'écriture avec doubles, triples,..., indices et sommes, écriture qui permet, d'une part, de visualiser une distance dans $\mathbb{R}^n$, et d'autre part, de ne pas retenir les nombres de modalités des différents facteurs.

Rappelons que dans l'exemple des forêts, $I = 3$ est le nombre de forêts et $n = 18$ le nombre total de données. Le test de l'hypothèse H_0 d'égalité des moyennes (de niveau α) se fait en comparant la valeur du rapport $\widehat{F}$ au quantile $F_{(I-1,n-I),1-\alpha}$ d'une loi de Fisher à $((I-1),(n-I))$ degrés de liberté. Pour rejeter l'hypothèse H_0, il faudra donc que $\widehat{F} > F_{(I-1,n-I),1-\alpha}$.

Remarque : la statistique F peut s'interpréter comme le rapport de la variabilité inter-groupe sur la variabilité intra-groupe. En effet le carré moyen du modèle

mesure l'écart des moyennes des groupes (forêts) à la moyenne générale ; c'est une mesure de variabilité entre les groupes (d'où la dénomination inter-groupe). Le carré moyen résiduel mesure l'écart de chaque individu (arbre) à la moyenne du groupe (forêt) auquel il appartient ; c'est une mesure de la variabilité à l'intérieur de chaque groupe (d'où la dénomination intra-groupe). C'est de ces conceptions que vient la dénomination "analyse de la variance".

2.3 Intervalle de confiance et test de Student

Si on choisit a priori (c'est-à-dire avant le résultat de l'expérience) deux populations à comparer, par exemple la 1 et la 2, on peut obtenir un intervalle de confiance pour la différence des moyennes théoriques $\mu_1 - \mu_2$. En effet, un estimateur "naturel" de cette différence est $\widehat{\mu}_1 - \widehat{\mu}_2 = Y_{1.} - Y_{2.}$ et on montre facilement que cet estimateur a pour variance

$$\sigma^2(\frac{1}{n_1} + \frac{1}{n_2}).$$

Par le même raisonnement que précédemment, notamment en utilisant le carré moyen résiduel $\widehat{\sigma^2} = \dfrac{1}{n-I} SC_1 = \dfrac{1}{n-I} \displaystyle\sum_{i,j} (Y_{i.} - Y_{..})^2$ comme estimateur de σ^2, on montre qu'un intervalle de confiance de niveau $1-\alpha$ pour $\mu_1 - \mu_2$ est donné par

$$\left[Y_{1.} - Y_{2.} - T_{n-I,1-\alpha/2} \sqrt{\widehat{\sigma^2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \ , \ Y_{1.} - Y_{2.} + T_{n-I,1-\alpha/2} \sqrt{\widehat{\sigma^2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right].$$

Comme précédemment, on en déduit également un test de Student de comparaison de ces 2 moyennes en regardant si la valeur zéro appartient à cet intervalle. Dans le cas présent, la statistique de test, $\widehat{T}$, est définie par

$$\widehat{T} = |Y_{1.} - Y_{2.}| \left(\widehat{\sigma^2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)^{-1/2}$$

et l'on comparera $\widehat{T}$ à $T_{n-I,1-\alpha/2}$ pour réaliser le test

$$\text{contre} \quad \begin{array}{rcl} H_0 & : & \mu_1 = \mu_2 \\ H_1 & : & \mu_1 \neq \mu_2 \end{array} .$$

Remarque : Ce test est différent (et sous nos hypothèses, plus puissant) que le test de comparaison de deux moyennes basé sur les deux seuls échantillons des groupes 1 et 2 car la variance résiduelle σ^2 est estimée sur l'ensemble des groupes.

3 Conclusion

Dans les deux problèmes évoqués dans ce chapitre, à savoir la régression linéaire simple et l'analyse de variance à un facteur, nous avons utilisé :

- le même type d'hypothèses sur les erreurs ;
- l'estimateur des moindres carrés ;
- des tests de Fisher et de Student.

En fait, ces deux problèmes ne sont pas si éloignés qu'ils le paraissent a priori car les deux modèles utilisés font partie d'une même famille de modèles : **le modèle linéaire statistique.**

4 Exemples traités par logiciels informatiques

Pour illustrer cette première approche du modèle linéaire, tout comme nous le ferons dans la plupart des chapitres suivants, nous présentons maintenant des exemples traités par logiciel informatique, avec la reproduction de l'essentiel des sorties informatiques post-traitements (sorties numériques et graphiques) nécessaire à leur étude. Nous allons utiliser sur les mêmes exemples les trois logiciels statistiques SAS, Splus et R. Pour les deux exemples de ce chapitre, qui comprennent un faible nombre de données, nous commençons par construire "à la main" les tables de données. Par la suite, nous nous permettrons de travailler directement sur données déjà construites. Nous rappelons que celles-ci sont disponibles sur le site :

`http://www.dunod.com`

4.1 Exemple de régression linéaire simple

Reprenons l'exemple de la tension artérielle en fonction de l'âge.

Logiciel SAS :

On commence par constituer la table `tension` que l'on écrit dans le répertoire `sasuser`. Compte tenu de la faible taille du jeu de données, nous avons utilisé la commande préhistorique `cards` comme "carte perforée".

```
data sasuser.tension;
input age tension;cards;
35 114
45 124
55 143
65 158
75 166; run;
```

Une fois ces commandes tapées dans l'éditeur et une fois soumises (soit directement soit en cliquant avec la souris sur `Run`, puis `Submit`. **Notons que cela devra être fait après chaque suites de commandes écrites dans l'éditeur**), on observe à l'écran la table `tension`. On peut alors maintenant écrire les commandes suivantes permettant la régression linéaire simple et qui s'appuient sur la procédure `reg` :

```
proc reg data=sasuser.tension; model tension=age;
plot tension*age; run; quit;
```

Interprétons maintenant cette seconde partie du programme :

- la première ligne (`proc reg ...`) déclare que la procédure va travailler sur la table "tension" contenue dans le répertoire `sasuser` (répertoire usuel de travail) ;
- la seconde ligne déclare la variable à expliquer (`tension`) et la variable explicative (`age`) ;
- la troisième ligne propose de tracer un graphique qui superpose le nuage de points avec la droite de regression. Ce graphique n'est utilisable qu'en regression linéaire simple. Dans le cas général, on consultera les exemples des chapitres qui suivent.

Le résultat de ces procédures, que l'on peut retrouver dans la fenêtre `Output` pour les résultats numériques, et dans une fenêtre `Graph` pour les graphiques (là-encore, cela est générique et devra être renouvelé après chaque suite de commandes tapées), est le suivant (on en présente ici un extrait) :

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	1	1904.40000	1904.40000	180.797	0.0009

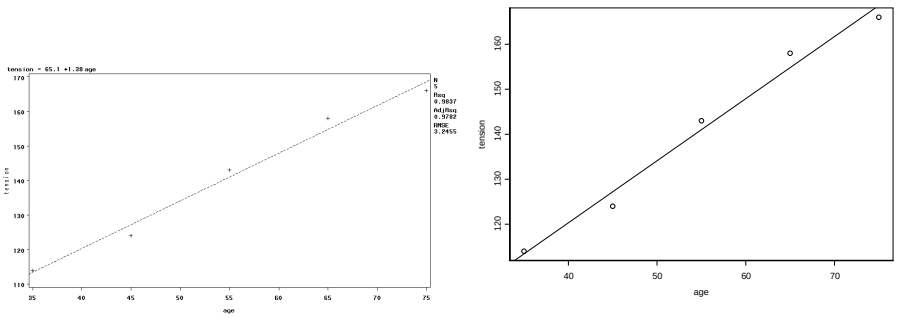


FIGURE 1.3 – Nuage de points et droite de régression linéaire associée en SAS (à gauche) et en Splus ou R (à droite)

```
Error             3     31.60000     10.53333
C Total           4   1936.00000

                          Parameter Estimates

                   Parameter       Standard    T for H0:
 Variable  DF       Estimate          Error   Parameter=0    Prob > |T|

 INTERCEP   1      65.100000     5.82837885        11.169        0.0015
 AGE        1       1.380000     0.10263203        13.446        0.0009
```

Le graphique 4.1 montre clairement que la dépendance entre les variables est approximativement linéaire. La droite de régression par moindres carrés est tracée et son équation apparaît en haut.

Commentons maintenant les autres résultats obtenus. Le tableau d'analyse de la variance indique essentiellement deux valeurs intéressantes : la première est la valeur indiquée par F, qui est égale (avec les notations des sections précédentes) à $\widehat{F} \simeq 180.797$. Elle correspond à la valeur de la statistique de Fisher testant l'hypothèse H_0 concernant la nullité de la pente β de la régression linéaire. Normalement, cette valeur $\widehat{F}$ est à comparer avec le quantile $F_{(1,3),1-\alpha}$ (non indiqué) où α est le niveau du test (souvent on choisit $\alpha = 5\%$). Cependant, le logiciel indique plutôt la P-value (notée $Pr > F$) qui donne la probabilité qu'une variable aléatoire suivant la loi considérée (ici la loi de Fisher avec les paramètres $(1, 3)$) puisse être supérieure à la valeur obtenue (ici 180.797).

D'une manière générale, on rejettera l'hypothèse H_0 lorsque la P-value sera inférieure à α, niveau du test (souvent on choisit $\alpha = 0.05$). On pourra se reporter au chapitre 2 pour obtenir plus de détails sur ces questions.

Sur cet exemple, la P-value est environ de 0.009. Si on avait choisi a priori un niveau $\alpha = 0.05$, on rejette l'hypothèse H_0 de la nullité de la pente de la regression. Cela donne une première légitimité à la modélisation des données par un modèle de régression linéaire simple.

On observe également les P-values associées aux statistiques de Student concernant l'hypothèse de nullité des deux paramètres (en reprenant les notations des sections précédentes, SAS appelle `INTERCEPT` le paramètre μ et `AGE` le coefficient multiplicateur de la variable `age`, soit β). Pour l'hypothèse de nullité de μ, $Prob > |T| \simeq 0.015$, pour celle de β, $Prob > |T| \simeq 0.009$: dans les deux cas, toujours en ayant choisit a priori $\alpha = 0.05$, on rejette donc les hypothèses de nullité de ces paramètres. Remarquons enfin que la deuxième P-value obtenue par un test de Student est de même valeur que celle obtenue par le test de Fisher : ceci n'est pas surprenant car nous avons évoqué le fait que ces deux tests, dans ce cadre, étaient strictement équivalents.

Logiciel Splus :

Nous allons maintenant reproduire les mêmes traitements avec le logiciel Splus. Pour ce qui concerne la régression linéaire, la commande principale est `lm`, mais une version plus conviviale et performante a été développée : `menuLm`. Voici donc le programme correspondant à ce qui a été fait en SAS :

```
age<-c(35,45,55,65,75)
tension<-c(114,124,143,158,166)
Tens<-data.frame(age,tension)

reg<-menuLm(tension~age,data=Tens)
plot(age,tension)
abline(reg)
```

Quelques commentaires sur ces commandes Splus :

- les 4 premières lignes servent à construire le tableau `Tens`. On commence par créer chaque variable, puis la commande `data.frame` permet de concaténer les deux variables pour former un tableau de données ;

- la commande `menuLm`, très riche en options, n'est utilisée ici que sous sa forme minimale ; sont ainsi présentés les résultats de la régression et la table d'analyse de la variance associée (voir ci-dessous) ;

- les deux dernières commandes permettent la construction du nuage de points et de la droite de régression linéaire associée.

Voici ce que l'on obtient (en partie) à l'écran :

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 65.1000  5.8284    11.1695  0.0015
        age  1.3800  0.1026    13.4461  0.0009

Residual standard error: 3.246 on 3 degrees of freedom
Multiple R-Squared: 0.9837
F-statistic: 180.8 on 1 and 3 degrees of freedom, the p-value is 0.0008894

Analysis of Variance Table         Response: tension

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
      age  1    1904.4 1904.400 180.7975 0.0008894091
Residuals  3      31.6   10.533
```

On retrouve le même type de résultats qu'avec SAS, à la nuance près des arrondis des différentes valeurs numériques.

Logiciel R :

Avec le logiciel R, on retrouve la plupart des commandes de Splus, à quelques différences de syntaxe près. En régression linéaire simple, la commande principale est `lm` :

```
age=c(35,45,55,65,75)
tension=c(114,124,143,158,166)
Tens=data.frame(age,tension)

reg=lm(tension~age,data=Tens)
summary(reg)
anova(reg)
```

```
plot(age,tension)
abline(reg)
```

Quelques commentaires sur ces commandes R :

- tout ce qui a été écrit en R serait lisible en Splus, en tenant compte du fait que la commande = doit être remplacée par la commande _ ou $<-$;
- la présentation des résultats doit être demandée et elle se fait toujours avec la commande `summary` (cela aurait été pareil si l'on avait utilisé la commande `lm` en Splus). La commande `anova` permet en plus la sortie de la table d'analyse de la variance (**attention de n'utiliser cette commande que dans le cas bien précis de la régression linéaire et non dans ceux de l'analyse de la variance ou de la covariance !**).

Voici un extrait des résultats de ces commandes :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.1000     5.8284   11.17  0.00154 **
age           1.3800     0.1026   13.45  0.00089 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.246 on 3 degrees of freedom
Multiple R-Squared: 0.9837,     Adjusted R-squared: 0.9782
F-statistic: 180.8 on 1 and 3 DF,  p-value: 0.0008894

Analysis of Variance Table                    Response: tension
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 1904.40 1904.40  180.80 0.0008894 ***
Residuals  3   31.60   10.53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On retrouve le même type de résultats numériques qu'avec SAS, et quasiment le même graphe qu'avec Splus (ce qui explique que cette figure n'est pas représentée). On peut remarquer cependant que l'on voit apparaître dans les résultats numériques obtenus avec R le niveau de signification des tests en plus de la P-value : cela nous semble assez sage (voir à ce propos "l'intermède métastatistique" au chapitre suivant 2).

4.2 Exemple de l'analyse de la variance à un facteur

Reprenons l'exemple des hauteurs d'arbres dans différentes forêts, pour lequel une analyse de la variance à un facteur va tester si cette hauteur est la même dans les différentes forêts que l'on a considérées. Pour continuer l'apprentissage des différents logiciels, nous considérons maintenant le cas où le fichier contenant les données est un fichier texte (de format ASCII) composé de 2 colonnes séparées par une tabulation, avec en tête de colonne le nom des variables correspondantes, c'est-à-dire `hauteur` et `foret`. Ce fichier est supposé être dans le répertoire `C:\Donnees`, et a pour nom `foret3.txt` (nom donné pour différencier le fichier du facteur `foret`. On aurait également pu associer une autre extension que `.txt` (ou pas d'extension) à ce nom de fichier.

Remarque : Dorénavant, pour ne pas surcharger la lecture des résultats, nous ne présenterons les résultats numériques et les graphiques que pour un seul logiciel (lorsque les autres n'apportent rien de plus).

Logiciel SAS :

L'exemple de l'analyse de la variance à un facteur obéit à la même logique que celui de la regression linéaire simple :

```
proc import out=sasuser.foret3 datafile="C:\Donnees\foret3.txt";
run;

proc glm data=sasuser.foret3;
class foret;
model hauteur=foret;
means foret;
output out=sortie r=residu p=predite;run;

proc gplot data=sortie;
plot residu*predite;run; quit;
```

Par rapport au premier programme présenté pour la régression linéaire, les petites différences sont :

- la procédure `import` va chercher les données dans le fichier `foret3.txt` (le nom des variables doit être en première ligne) et les transforme en une table SAS enregistrée dans le répertoire de travail `sasuser`. Remarquons que pour avoir des graphiques de résidus plus réalistes, nous n'avons pas utilisé les données

présentées précédemment, mais un jeu plus important (soit 37 données) extrait du livre Dacunha-Castelle & Duflo [20]. Il est également à noter que cette étape d'importation (ou d'exportation) de données se fait de façon beaucoup plus agréable sur les dernières versions de SAS, avec la souris après avoir "cliqué" dans `File`, puis sur `Import Data` (ou `Export Data`).

– pour l'étape de procédure principale, on a dû bien sûr utiliser une procédure adaptée à l'analyse de variance. Par souci de cohérence avec ce qui suit, nous avons préféré utiliser la procédure `glm` plutôt que la procédure `anova`.

– la ligne commençant par `class` déclare que la variable `foret` (le numéro de la forêt) est qualitative.

– la ligne commençant par `means` demande explicitement les moyennes qui ne sont pas données par défaut.

– la ligne commençant par `output` réalise une sortie des résultats de l'analyse : les résidus et les valeurs prédites présentés sur la liste `sortie`. Ces données sont reprises par `gplot` qui donne un graphique haute résolution.

Voici un extrait des résultats obtenus :

```
Dependent Variable: hauteur
                                      Sum of
Source                        DF         Squares     Mean Square    F Value    Pr > F
Model                          2      49.8483020      24.9241510       7.20    0.0025
Error                         34     117.6273736       3.4596286
Corrected Total               36     167.4756757

                      R-Square     Coeff Var      Root MSE    hauteur Mean
                      0.297645      7.443249      1.860008        24.98919

Source                     DF       Type I SS     Mean Square    F Value    Pr > F
foret                       2     49.84830205     24.92415102       7.20    0.0025

Source                     DF     Type III SS     Mean Square    F Value    Pr > F
foret                       2     49.84830205     24.92415102       7.20    0.0025

                      Level of            ------------hauteur------------
                      foret         N             Mean          Std Dev
```

```
1          13          25.9923077          1.39072016
2          14          25.3857143          1.77324437
3          10          23.1300000          2.43905719
```

Le graphique des résidus (voir figure 4.2) ne met pas en évidence d'écart manifeste aux hypothèses faites sur le bruit ε du modèle (les résidus sont à peu près répartis de la même manière pour chaque modalité du facteur). Le tableau d'analyse de la variance (**en général, on ne considérera que les résultats issues de l'analyse dite de type "III" ; cependant, dans le cas d'un seul facteur, les résultats de l'analyse dite de type "I" sont identiques** ; on se reportera au chapitre 5 pour plus de détails) montre que les trois forêts ont des hauteurs significativement différentes : en effet on obtient que $\widehat{F} \simeq 7.20$, valeur qui est à comparer avec le quantile $F_{(2,34),0.95}$ (non indiqué) et dont la P-value associée vaut $\simeq 0.0025$ (ce qui est inférieur à 0.05) : avec $\alpha = 0.05$, on rejette donc l'hypothèse H_0 que les forêts ont la même hauteur. Dans ce jeu de données, le facteur `foret` est donc influent sur la hauteur d'une forêt. Observons enfin les différentes valeurs obtenues par la commande `means` : les hauteurs estimées pour chacune des forêts valent approximativement et respectivement 25.99, 25.39 et 23.13.

Logiciel Splus :

Sur le même exemple des hauteurs de forêts, l'analyse de la variance par Splus peut s'écrire de la manière suivante :

```
import.data(DataFrame="foret3",FileName="C:/Donnees/foret3.txt",FileType="ASCII")
foret3$foret<-as.factor(foret3$foret)

menuAov(hauteur~foret,foret3,plotResidVsFit.p=T)
menuAov(hauteur~foret-1,foret3,coef.p=T)
```

Commentaires :

- les deux premières lignes de commande permettent de définir proprement le `data.frame` `foret3` à partir d'un fichier `foret3.txt` qui se présente comme un fichier ASCII (ici un fichier texte dans lequel on a écrit les noms des variables et les valeurs prises sous forme de deux colonnes). La transformation de la variable `foret` en un objet `factor` est nécessaire pour réaliser une analyse de la variance.

- la commande `menuAov` permet d'effectuer des analyses de la variance et contient un très grand nombre d'options. Ici, elle est utilisée deux fois. En effet, pour

obtenir la table d'analyse de la variance avec les résultats des tests de Fisher associés, il faut préciser le modèle "naturel" : `hauteur foret` qui correspond à une analyse de la variance de la variable `hauteur` à partir du facteur `foret`. De plus, le tracé du graphes des résidus en fonction des valeurs prédites est demandé par l'option `plotResidVsFit.p=T` (nous n'avons choisi qu'un seul graphes, mais beaucoup d'autres auraient été disponibles également). Cependant, **cette première commande `menuAov` ne permettra pas d'obtenir les estimations des différents coefficients du modèle**. Pour ce faire, on devra poser un **modèle sans `intercept` soit : `hauteur foret-1` dans cette seconde commande `menuAov`.** On précisera ensuite l'option `coef.p`, qui fournit les moyennes pour chaque classe du facteur. **Il ne faudra alors pas tenir compte de la table d'analyse de la variance pour cette seconde commande `menuAov`**. Pour plus de détails sur ces questions, voir la partie suivante 4.3.

On obtient ainsi les résultats suivants, qui semblent, en ce qui concerne la seule analyse de la variance (et c'est ce qui nous importe le plus) assez proches dans leur contenu et dans leur présentation avec ceux obtenus avec SAS :

```
>menuAov(hauteur~foret,foret3)

                  foret Residuals
 Sum of Squares  49.8483  117.6274
Deg. of Freedom        2        34

Residual standard error: 1.860008 Estimated effects may be
unbalanced

          Df Sum of Sq  Mean Sq  F Value       Pr(F)
    foret  2   49.8483 24.92415 7.204285 0.002462993
Residuals 34  117.6274  3.45963


>menuAov(hauteur~foret-1,foret3,coef.p=T,plotResidVsFit.p=T)

                  foret Residuals
 Sum of Squares 23154.85    117.63
Deg. of Freedom        3        34

Residual standard error: 1.860008
Estimated effects are balanced
```

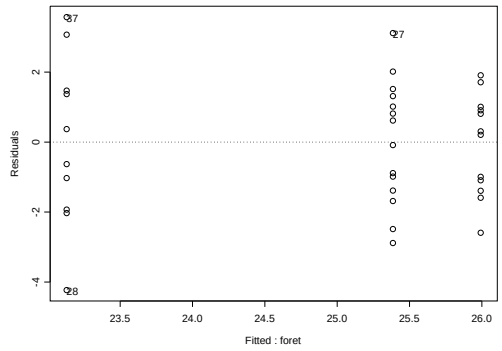


FIGURE 1.4 – Graphique des résidus en fonction des valeurs prédites (Splus)

```
          Df Sum of Sq  Mean Sq  F Value Pr(F)
    foret  3  23154.85 7718.284 2230.957     0
Residuals 34    117.63    3.460

Estimated Coefficients:
   foret1   foret2 foret3
 25.99231 25.38571  23.13
```

Remarquons juste que le graphique des résidus en fonction des valeurs prédites (Figure 4.2) fournit les numéros des 3 individus ayant les valeurs absolues de résidus maximales : cela peut être utile pour mettre en évidence et éventuellement éliminer des "outliers".

Logiciel R :

Voyons maintenant les commandes nécessaires en R pour obtenir un traitement équivalent :

```
foret3=read.table("C:/Donnees/foret3.txt",header=TRUE)
foret3$foret=as.factor(foret3$foret)

lm.foret=lm(hauteur~foret,foret3)
library(car)
Anova(lm.foret,type="III")
```

```
par(mfrow=c(2,2))
plot(lm.foret,las=1)

coeff.foret=lm(hauteur~foret-1,foret3)
summary(coeff.foret)
```

Commentaires :

- les deux premières lignes reprennent ce qui a été fait en Splus, à la nuance près de l'option `header=TRUE` qui permet de prendre en compte le nom des variables déjà indiqué dans le fichier texte `foret3.txt`.

- pour effectuer l'analyse de la variance de `hauteur` par le facteur `foret`, on utilise la commande `lm` avec le modèle "naturel" : `hauteur`∼`foret`. Ensuite, **pour pouvoir effectuer une analyse de la variance suffisamment générale (voir plus loin), on fait appel à une commande `Anova` qui n'existe pas dans le module de base de R, mais dans une librairie ("package") appelée `car`.** Les deux commandes qui suivent permettent le tracé sur le même graphe des différentes représentations graphiques qui s'avéreront être utiles : résidus (standardisés ou non) en fonction des valeurs prédites, le QQ-plot des résidus et enfin la distance de Cook des résidus (ce qui mesure l'influence de la mesure sur le paramètre estimé) en fonction du numéro des différentes observations (nous donnerons plus de détails sur ces graphes dans le chapitre 4).

- enfin, tout comme avec le logiciel Splus, on se heurte à la difficulté de l'estimation des coefficients estimés à partir de ce premier modèle. **Pour retrouver les valeurs des coefficients $\widehat{\mu}_1$, $\widehat{\mu}_2$ et $\widehat{\mu}_3$, on usera de la même "astuce" qu'avec Splus, c'est-à-dire que l'on précisera le modèle sans `intercept` par : `hauteur`∼`foret-1` ;** d'où les deux dernières commandes.

Voici les résultats numériques :

```
> summary(lm.foret)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.9923     0.5159  50.385  < 2e-16 ***
foret2       -0.6066     0.7164  -0.847 0.403075
foret3       -2.8623     0.7824  -3.659 0.000851 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.86 on 34 degrees of freedom
Multiple R-Squared: 0.2976,     Adjusted R-squared: 0.2563
F-statistic: 7.204 on 2 and 34 DF,  p-value: 0.002463


> Anova(lm.foret,type="III")
Anova Table (Type III tests)

Response: hauteur
            Sum Sq Df   F value    Pr(>F)
(Intercept) 8782.8  1 2538.6542 < 2.2e-16 ***
foret         49.8  2    7.2043  0.002463 **
Residuals    117.6 34
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> summary(coeff.foret)

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
foret1  25.9923     0.5159   50.38   <2e-16 ***
foret2  25.3857     0.4971   51.07   <2e-16 ***
foret3  23.1300     0.5882   39.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.86 on 34 degrees of freedom
Multiple R-Squared: 0.9949,     Adjusted R-squared: 0.9945
F-statistic:  2231 on 3 and 34 DF,  p-value: < 2.2e-16
```

Le graphe des résidus est quasiment identique à la figure 4.2, les numéros des 3 individus ayant les valeurs absolues de résidus maximales sont également mis en évidence. On s'aperçoit ici des valeurs différentes obtenues par la statistique de Fisher suivant le modèle précisé (7.204 et 2 degrés de liberté sur 34 pour la premier et bon modèle, 2231 et 3 degrés de liberté sur 34 pour le second modèle, qui rappelons le, n'est présent que pour l'obtention des paramètres estimés).

Commentaire général : Ces analyses montrent une bonne adéquation du modèle d'analyse de la variance. Ce modèle permet de conclure à une différence significa-

tive entre les hauteurs des trois forêts, les hauteurs estimées étant respectivement : $\widehat{\mu}_1 \simeq 25.99$, $\widehat{\mu}_2 \simeq 25.39$ et $\widehat{\mu}_3 \simeq 23.13$.

4.3 Ce qu'il ne faut pas faire en analyse de la variance pour estimer les paramètres...

On reprend les données et certaines commandes de l'exemple informatique précédent. Voyons maintenant certains pièges dans lesquels on peut très facilement tomber.

Logiciel SAS :

Il ne faut surtout pas remplacer les commandes précédentes par la suite de commandes suivantes :

```
proc glm data=sasuser.foret3;
class foret;
model hauteur=foret/solution; run;
```

Voici un extrait de la sortie, là où elle diffère de celle de l'exemple précédent :

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		23.13000000	B	0.58818608	39.32	<.0001
foret	1	2.86230769	B	0.78236062	3.66	0.0009
foret	2	2.25571429	B	0.77011635	2.93	0.0060
foret	3	0.00000000	B	.	.	.

La commande `solution`, très, trop souvent utilisée, amène des estimations des paramètres qui diffèrent de celles obtenues lors de l'analyse de la variance précédente. Des explications quant à ces nouvelles estimations seront données ci-dessous, mais retenons d'ores et déjà qu'il faut utiliser la commande `solution` avec beaucoup, beaucoup de prudence.

Logiciel Splus :

Voici, ce que peut-être naturellement, on aurait été tenté de demander au logiciel pour effectuer la même analyse de la variance, et que l'**on doit éviter de faire :**

```
menuAov(hauteur~foret,foret3,means=T,lsmeans=T,coef.p=T)
```

Voici une partie des résultats obtenus :

```
Estimated Coefficients:
 (Intercept)     foret1     foret2
    24.83601 -0.3032967 -0.8530037

Tables of means
Grand mean

 24.989

 foret
         1      2     3
    25.992 25.386 23.13
rep 13.000 14.000 10.00

Tables of adjusted means
Grand mean

   24.83601
se  0.30898

 foret
        1      2      3
   25.992 25.386 23.130
se  0.516  0.497  0.588
Warning messages:
  Design is unbalanced. Standard errors are not computed for type = "mea
ns" in: model.tables.aov(object, type = "means", se = T)
```

En plus de la méfiance à accorder aux différentes estimations des coefficients (voir plus haut et plus bas), remarquons que les résultats obtenus avec la commande `means` (table des moyennes, notée `Tables of means`) permettent de visualiser les estimations $\widehat{\mu}_1$, $\widehat{\mu}_2$ et $\widehat{\mu}_3$), les effectifs associés (un message précise que les écart-types ne peuvent être évalués par cette commande du fait que les effectifs ne sont pas équilibrés), et ce qui est appelé `Grand mean`, c'est-à-dire la moyenne arithmétique des `hauteur`. Les résultats relatifs à la commande `lsmeans` (moyenne ajustée) se trouvent après ce qui est appelé `Tables of adjusted means`, et propose ici des résultats que nous n'utiliserons par : la valeur dite `Grand mean` calcule la moyenne arithmétique des coefficients

estimés (donc $\frac{1}{3} \cdot (\widehat{\mu}_1 + \widehat{\mu}_2 + \widehat{\mu}_3)$) sans tenir compte des effectifs. De plus, **les "standard errors" (notées `se`) calculées pour chaque coefficient estimé ne sont pas issues de la définition usuelle, et ne sont pas par nous utilisables.**

Logiciel R :

Voici, de la même manière, ce que, presque naturellement, on écrirait pour effectuer une analyse de la variance et estimer les coefficients, mais qu'il vaut mieux éviter :

```
lm.foret=lm(hauteur~foret,foret3)
lm.foret$coef
```

Les résultats (extraits) de ces commandes sont :

```
(Intercept)      foret2      foret3
 25.9923077  -0.6065934  -2.8623077
```

Comme nous l'avons vu précédemment, la commande `lm` permet donc d'effectuer une analyse de la variance aussi bien qu'une regression. On aurait également pu utiliser la commande `aov` dont les actions sont similaires, mais qui affiche un peu moins de détails. La suite est plus gênante, puisqu'on demande d'afficher les coefficients estimés (par la commande `lm.foret$coef`) et comme nous allons le préciser maintenant, on n'obtient pas vraiment ces coefficients...

Commentaire général sur la partie 4.3 : Attention ! Il faut faire très attention avec les différentes estimations des coefficients du modèle et de leurs écart-types telles qu'elles sont proposées sur ces différentes pages de résultats. Les différents coefficients estimés ne correspondent pas aux coefficients $\widehat{\mu}_1$, $\widehat{\mu}_2$ et $\widehat{\mu}_3$ définis dans le traitement théorique de cet exemple. Par exemple, avec le logiciel SAS, le coefficient appelé `intercept` correspond à $\widehat{\mu}_3$, celui appelé `foret1` est $\widehat{\mu}_1 - \widehat{\mu}_3$, celui appelé `foret2` est $\widehat{\mu}_2 - \widehat{\mu}_3$ et enfin celui appelé `foret3` est $\widehat{\mu}_3 - \widehat{\mu}_3 = 0$. Avec le logiciel R, c'est le coefficient $\widehat{\mu}_1$ qui devient la référence, et alors ce qui est appelé `Intercept` est $\widehat{\mu}_1$, ce qui est appelé `foret2` est $\widehat{\mu}_2 - \widehat{\mu}_1$ et enfin celui appelé `foret3` est $\widehat{\mu}_3 - \widehat{\mu}_1$. Pour complexifier encore cet imbroglio, avec le logiciel Splus, ce qui est appelé `Intercept` est $\frac{1}{3} \cdot (\widehat{\mu}_1 + \widehat{\mu}_2 + \widehat{\mu}_3)$, et ... on ne comprend plus du tout ce qui est appelé `foret1` et `foret2`. Attention donc à utiliser exclusivement les commandes évoquées dans la sous-section précédente 4.2.

5 Exercices

Exercice 1.1

(*) Soit $y_1, \cdots, y_n$ des réels. Déterminer le réel $\widehat{m}$ qui minimise $SCR(m) = \sum_{i=1}^{n} |y_i - m|^2$ (Indication : on peut dériver la fonction $m \mapsto SCR(m)$). Que représente $\widehat{m}$ par rapport à $y_1, \cdots, y_n$?

Exercice 1.2

(**) Soit $y_1, \cdots, y_n$ des réels. Déterminer le réel $\tilde{m}$ qui minimise la somme des valeurs absolues $SVA(m) = \sum_{i=1}^{n} |y_i - m|$ (on pourra traiter d'abord le cas où $n = 2p + 1$ et classer les réels (y_i) sous la forme $y_{(1)} \leq y_{(1)} \leq \ldots \leq y_{(n-1)} \leq y_{(n)}$, puis passer au cas où $n = 2p$). Que représente $\tilde{m}$ par rapport à $y_1, \cdots, y_n$?

Exercice 1.3

(*) Soit $(x_1, y_1), \cdots, (x_n, y_n)$ des couples de réels. Déterminer le réel $\widehat{a}$ qui minimise la somme des carrés résiduelle $SCR(a) = \sum_{i=1}^{n} |y_i - a \cdot x_i|^2$ (Indication : on peut dériver la fonction $a \mapsto SCR(a)$). Que représente $\widehat{a}$ par rapport à $y_1, \cdots, y_n$? Comparer avec la valeur $\widehat{\beta}$ obtenue dans le cadre de la régression linéaire simple classique.

Exercice 1.4

(**) [Maximum de vraisemblance] On rappelle que, dans le cadre d'un modèle statistique paramétrique (c'est-à-dire que la loi du modèle ne dépend que d'un nombre fini de paramètres), la densité des observations $Y_1, \ldots, Y_n$ vue comme une fonction des paramètres est appelé la vraisemblance. L'estimateur du maximum de vraisemblance de ces paramètres est la valeur des paramètres qui maximise la vraisemblance.

i. Considérons le modèle gaussien (1.1) et ses paramètres (μ, β, σ^2). Montrer que $-2\times$ log-vraisemblance vaut

$$L(\mu, \beta, \sigma^2) = n \cdot log(2\pi) + n \cdot log(z) + \frac{1}{z} \cdot \sum_{i=1}^{n} (Y_i - \mu - \beta Z_i)^2,$$

en notant $z = \sigma^2$ pour dériver plus facilement.

ii. Comparer les estimateurs du maximum de vraisemblance avec ceux des moindres carrés de μ et β.

iii. Comparer (au sens de la vitesse de convergence) l'estimateur du maximum de vraisemblance de σ^2 avec $\widehat{\sigma}^2$ défini dans le corps du chapitre.

iv. Mêmes questions pour les estimateurs du maximum de vraisemblance de $(\mu_1, \cdots, \mu_I, \sigma^2)$ dans le cadre du modèle gaussien d'analyse de la variance (1.6).

Exercice 1.5

(***) [Test du rapport de vraisemblance] On rappelle que dans le cadre de variables aléatoires $Y_1, \ldots, Y_n$ suivant un modèle statistique et admettant une densité, la statistique U du rapport de vraisemblance est le rapport entre le maximum (en les paramètres) de la vraisemblance sous l'hypothèse H_1 et le maximum (en les paramètres) de la vraisemblance sous l'hypothèse H_0. La région d'acceptation du test est du type $U \leq \lambda$, avec $\lambda \in \mathbb{R}$ dépendant du niveau du test.

i. On considère le test du rapport de vraisemblance de $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$ dans le cadre du modèle gaussien (1.1). Montrer que si l'on pose

$$SC := \sum_{i=1}^{n} (Y_i - \widehat{\mu} - \widehat{\beta} Z_i)^2,$$

alors le maximum sous H_1 de la log-vraisemblance vaut

$$-\frac{1}{2}\big(n \cdot \log(2\pi) + n \cdot \log(SC/n) + n\big)$$

ii. Montrer un résultat équivalent sous H_0 en posant

$$\widetilde{SC} := \sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2,$$

iii. Donner l'expression de $\log(U)$ et comparer avec les tests de Student et Fisher.

iv. Déterminer le test du rapport de vraisemblance pour le test $H_0 : \mu_1 = \mu_2 = \mu_3$ contre H_1 : on n'a pas $\mu_1 = \mu_2 = \mu_3$, dans le cadre du modèle gaussien (1.6). Comparer avec le test d'analyse de la variance de Fisher.

Chapitre 2

Intermède métastatistique : pour une théorie pratique des tests

Dans ce chapitre, nous présentons les méthodes de construction de tests dans les cas les plus courants. Nous définissons la notion de P-value et nous tentons de répondre à la question : comment fixer le niveau d'un test ?

1 Erreurs associées à un test

Un test a le plus souvent pour but de choisir entre deux hypothèses :

- l'une, dite "l'hypothèse générale", notée H_1, qui correspond au modèle dans sa pleine généralité ;
- l'autre, dite "l'hypothèse nulle", notée H_0, qui correspond à un certain sous-modèle défini, par exemple, par la nullité de certains paramètres (d'où son nom). Cette hypothèse s'inscrit souvent dans la démarche scientifique qui consiste à privilégier les modèles simples.

Un test est une fonction des observations qui choisit entre ces deux hypothèses. Le résultat du test est donc H_0 ou H_1 . Notons que **rejeter l'hypothèse H_0 est beaucoup riche en information que l'accepter.** C'est pour cela que dans ce dernier cas on dit que le test est significatif .

Dans la plupart des cas, et, en fait, dans tous les cas considérés dans cet ouvrage, la construction d'un test est fondée sur une statistique, c'est-à-dire une fonction (mesurable) des observations, qui a deux propriétés essentielles :

- sa loi de probabilité $\mathcal{L}$ est connue sous H_0 (c'est-à-dire lorsque H_0 est vraie) et elle ne dépend pas des paramètres restant variables. On dit que c'est une *statistique libre*.

- sa loi sous H_1 a tendance à prendre soit (a) des très grandes valeurs, (b) des toutes petites valeurs, (c) soit les deux.

Par exemple, dans le modèle de régression linéaire simple, la statistique $\widehat{T}$ du test de Student de l'hypothèse "$\beta = 0$" donné par la formule (1.5)

- suit une loi de Student à $n-2$ degrés de libertés sous l'hypothèse nulle $\beta = 0$. Elle ne dépend pas des deux paramètres restant variables : μ et σ^2.

- a tendance à tendre vers $+\infty$ si $\beta > 0$ et vers $-\infty$ si $\beta < 0$. Nous sommes donc dans la situation (c).

On définit l'*erreur de première espèce* ou *niveau* α du test comme la probabilité de se tromper (c'est-à-dire de choisir H_1) si H_0 est vraie :

$$\alpha = \mathbb{P}\,(\text{"}H_0 \text{ est rejetée par le test"} \mid \text{"}H_0 \text{ est vraie"}).$$

Suivant la situation (a), (b) ou (c), on obtient un test de niveau α en définissant la région de rejet :

(a) *unilatérale droite* : on rejette H_0 si $\widehat{T} > L_{1-\alpha}$. où $L_{1-\alpha}$ est le fractile d'ordre $1-\alpha$ de la loi $\mathcal{L}$.

(b) *unilatérale gauche* : cas symétrique du précédent, la région de rejet est maintenant $]-\infty, L_\alpha[$.

(c) *bilatérale* : dans ce cas, la loi $\mathcal{L}$ est le plus souvent symétrique auquel cas on rejette H_0 si $|\widehat{T}| > \mathcal{L}_{1-\alpha/2}$

(pour la définition des fractiles ou quantiles voir en appendice). Par construction ces tests sont de niveau α.

On définit également l'*erreur de seconde espèce* ou *manque de puissance* β comme

$$\beta = \mathbb{P}\,(\text{"}H_0 \text{ n'est pas rejetée par le test"} \mid \text{"}H_1 \text{ est vraie"}).$$

Malheureusement cette quantité dépend fortement de l'écart à l'hypothèse nulle : il est facile de mettre en évidence une grande différence et difficile de mettre en évidence une petite. On ne sait donc pas en général calculer la puissance et on doit se contenter de résultats théoriques (voir chapitre 3) qui montrent que dans certains cas elle est optimale.

En résumé, **on ne construit pas un test mais une famille de tests dépendant de α. Comment alors fixer ce niveau α ?** Voila une question cruciale sur laquelle les manuels de statistique sont cruellement muets. Les logiciels de statistique ne sont pas d'un grande aide non plus, dans le mesure où, sauf dans des cas complexes, ils évitent soigneusement de trancher ce débat et donnent le niveau de signification (appelé également la "P-value") du (ou de la famille de) test(s), ce qui permet de transférer la responsabilité du choix de α à l'utilisateur.

2 La P-value

La P-value ou niveau de signification est en fait la valeur critique de α qui fait basculer le résultat du test. Dans tous les exemples cités elle est unique et on a

$$\text{si} \quad \alpha > P - \text{value}, \quad \text{on rejette } H_0; \tag{2.1}$$
$$\text{si} \quad \alpha < P - \text{value}, \quad \text{on ne rejette pas } H_0. \tag{2.2}$$

Prenons le cas d'une région de rejet unilatérale droite, cas (a), et soit $\widehat{t}$ la valeur de la statistique $\widehat{T}$ calculée sur les données (c'est-à-dire que $\widehat{t}$ est une réalisation de la variable aléatoire $\widehat{T}$). Dans ce cas, la P-value peut être explicitement calculée :

$$P - \text{value} = \mathbb{P}\,\{\widehat{T} > \widehat{t} \mid H_0\}.$$

Cela permet d'évaluer si la valeur $\widehat{t}$ est "vraisemblable" pour la loi $\mathcal{L}$ de $\widehat{T}$ sous l'hypothèse nulle H_0. Des calculs analogues sont possibles dans les cas (b) et (c), nous les laissons à titre d'exercice au lecteur. Une autre façon de considérer la P-value est de remarquer que sa loi sous H_0 est en fait une loi uniforme $[0, 1]$ et que donc passer de $\widehat{T}$ à P-value revient à transformer la loi $\mathcal{L}$ en une loi uniforme sur $[0, 1]$.

En pratique, il reste toujours à fixer α.

3 L'erreur de troisième espèce

D'après l'historien de la statistique S. Stigler, l'absence de références précises dans ce choix du α aurait été une gène pour des statisticiens du XVIIIème siècle comme Bernoulli ou Bayes. En fait les valeurs de α utilisées dans la littérature s'échelonnent de 20% à 0.1% (rappelons juste que plus α est proche de 0, plus on accepte facilement H_0). Par exemple si un effet n'est pas significatif à 20%, c'est-à-dire que P-value > 0.2, tous les statisticiens du monde seront d'accord pour déclarer que les données sont totalement "plates". Réciproquement, si un effet est significatif à 0.1%, ces mêmes statisticiens seront d'accord pour déclarer l'expérience très significative. Il n'en reste pas moins que la valeur critique qui fera basculer le plus souvent la conclusion de "plutôt négatif mais légère présomption qui mériterait une autre expérience" à "expérience positive mais qui nécessite confirmation", est le niveau $\alpha = 5\%$. En particulier si vous avez suivi des cours de statistiques, vous avez remarqué qu'au moment de réaliser l'application numérique, on vous a suggéré, comme par hasard, et sans aucune justification, d'utiliser $\alpha = 5\%$. Quel est l'origine de ce choix ?

A première vue, une telle valeur peut paraître élevée : par exemple, en contrôle de qualité, cela correspond à une fausse alerte sur 20 cas. En fait, une telle proportion est supportable en raison de ce que l'on pourrait appeler

"L'erreur de 3-ème espèce".

Cette notion est une boutade qui souligne un fait bien connu de toute personne qui a travaillé sur les résultats d'expériences réelles : toute "manip", même conduite dans les meilleures conditions, peut donner lieu à des divergences inexpliquées. C'est ainsi que l'on peut trouver par exemple dans une expérience sur l'étude de désinfectants, plus de germes sur les échantillons désinfectés que sur les échantillons témoins non désinfectés... On peut également mesurer parfois des quantités dont tous les experts savent qu'elles ne sont pas réalistes. Les causes de telles incertitudes sont souvent multiples et incluent même les interversions de variables et de fichiers informatiques.

L'"erreur de 3ème espèce" est donc la probabilité de travailler sur des données qui n'ont pas le sens qu'on voulait bien leur prêter (par une certaine modélisation par exemple). Elle est très difficilement quantifiable et elle doit être combattue par tout moyen graphique dont on dispose, mais, malheureusement, elle est toujours très loin d'être nulle. Pour cette raison, une expérience isolée, même très significative, n'est pas considérée comme une preuve suffisante. Pour cette raison également, un niveau de 5% n'est pas trop élevé : si on répète l'expérience dans des conditions indépendantes, la fausse alerte n'aura pas lieu plusieurs fois.

4 La canonisation du 5%

Une question reste cependant en suspend : comment les statisticiens ont-ils convergé vers cette habitude de choisir $\alpha = 5\%$, qui apparaît finalement comme une norme officieuse ?

Un premier élément de réponse nous pousse à nous retourner vers un des pères fondateurs de la statistique, Ronald Fisher, que l'on pourrait être tenté de désigner comme le responsable de cette "tradition" du 5%. En effet, si on considère dans son premier livre "Statistical methods for research workers" publié en 1925 [27], les différents exemples proposés, on trouve :

- dans l'exemple 20 (comparaison de deux médicaments), un niveau de signification qui est entre 0.1 et 0.05, ce qui n'est pas jugé significatif : "The value of p (il s'agit de la P-value) is, therefore, between 0.1 and 0.05 and cannot be regarded as signifiant".

- dans l'exemple 22 en revanche, la P-value est maintenant entre 0.05 et 0.02 et Fisher écrit : "the result must be judged as significant, though barely so".

En second lieu, l'histoire veut également que Karl Pearson, qui n'était pas en excellents termes avec Ronald Fisher, n'ait pas permis à ce dernier de reproduire les tables de Biometrika. Ainsi, pour la première édition (1925) de "Statistical methods for research workers", Fisher, qui avait dû recalculer les tables du test qui allait devenir épônyme, s'était limité, faute d'énergie, à 5%.

Enfin, on peut avancer des raisons plus "utilitaires" de cet emploi du 5%. Tout d'abord, dans un test bilatéral sur la moyenne, après emploi du théorème de la limite centrale, le quantile de la loi normale centrée réduite vaut 1.96, c'est-à-dire une valeur proche de 2 qui sera facilement utilisée (ce qui avait un intérêt réel en l'absence de moyens informatiques). Autre piste, 20 représente le nombre de doigts d'un humain, le premier grand nombre qui lui est naturel. Aussi une valeur inférieur à 1/20 peut lui apparaître comme le premier nombre négligeable...

S. Stigler appelle la "canonisation du 5%" la convergence tout autour de la planète vers ce consensus arbitraire qui a joué un rôle important dans le développement de la statistique. En effet, dans la statistique mathématique théorique telle qu'elle est présentée classiquement, rien n'est dit sur le choix du α, chaque statisticien ayant son α personnel qui est **subjectif**. Il n'y a pas d'universalité de la démarche statistique, même en ce qui concerne la statistique mathématique la plus "canonique", dès qu'elle en vient à être appliquée. Et pourtant, c'est également cette absence d'universalité qui

est reprochée à la statistique bayésienne alors que cette dernière a tant d'avantages par ailleurs. Le passage de la théorie aux données, et, plus généralement et en termes plus kantiens, celui de la réalité en soi au monde des phénomènes, ne peut s'appuyer *in fine* que sur des consensus, logiquement cohérents et non excessifs, dont l'origine est essentiellement sociale.

Ainsi, dans le cas des tests statistiques, une certaine universalité a été restaurée à partir du moment où tous les statisticiens de la planète ont réussi à se mettre d'accord sur une valeur commune, même arbitraire, du niveau α.

Finalement, nous dirons que le seul argument "objectif" solide en faveur du 5% consiste à dire que cette valeur est utilisée de manière régulière depuis plus de 50 ans et qu'elle n'a pas trop mal convenu.

Chapitre 3

Introduction au modèle linéaire statistique

Dans ce chapitre, nous définissons de manière générale mais sous certaines hypothèses (postulats) le modèle linéaire et nous donnons les formules et propriétés essentielles.

1 Écriture matricielle de modèles simples

Dans cette partie, nous donnons une présentation unifiée des différents exemples de modèles de la partie précédente et nous présentons de nouveaux modèles.

1.1 Régression linéaire simple

Nous reprenons le modèle de régression linéaire simple (1.1) du chapitre précédent,

$$Y_i = \mu + \beta \cdot Z_i + \varepsilon_i \quad i = 1, \cdots, n = 5.$$

On considère les vecteurs $Y = (Y_i)_{1 \leq i \leq 5}$ et $\varepsilon = (\varepsilon_i)_{1 \leq i \leq 5}$. Le modèle (1.1) s'écrit alors :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{pmatrix} = \begin{pmatrix} 1 & Z_1 \\ 1 & Z_2 \\ 1 & Z_3 \\ 1 & Z_4 \\ 1 & Z_5 \end{pmatrix} \begin{pmatrix} \mu \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$$

ou encore

$$Y = X \cdot \theta + \varepsilon \quad \text{avec} \quad \theta = \begin{pmatrix} \mu \\ \beta \end{pmatrix} \quad \text{et} \quad X = \begin{pmatrix} 1 & Z_1 \\ 1 & Z_2 \\ 1 & Z_3 \\ 1 & Z_4 \\ 1 & Z_5 \end{pmatrix}. \tag{3.1}$$

Attention ! pour conserver les notations usuelles, nous utiliserons la même typographie pour les matrices et les vecteurs, à savoir une lettre majuscule. Cependant, X sera en général une matrice, quand Y et Z seront des vecteurs.

1.2 Analyse de la variance à un facteur

On reprend l'exemple des trois forêts avec le modèle (1.6), et de la même manière que précédemment, on utilise la notation matricielle :

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{25} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{34} \\ Y_{35} \\ Y_{36} \\ Y_{37} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{25} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \\ \varepsilon_{34} \\ \varepsilon_{35} \\ \varepsilon_{36} \\ \varepsilon_{37} \end{pmatrix}$$

ce qui s'écrit encore

$$Y = X \cdot \theta + \varepsilon \quad \text{avec} \quad \theta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}. \tag{3.2}$$

Remarque : Dans l'exemple ci-dessus, Y a ses éléments indicés par deux indices i et j. Pourtant nous avons appelé, et nous appellerons encore Y "vecteur". En toute

rigueur un vecteur est simplement un membre d'un espace vectoriel, c'est-à-dire un espace stable par addition et multiplication scalaire. Il n'est pas exigé qu'il soit sous forme "vecteur colonne". C'est ainsi que l'on peut parler de l'espace "vectoriel des matrices" et voir donc une matrice, tableau à double entrée, comme un vecteur. Ce n'est que lorsque l'on veut faire du calcul explicite et écrire des produits matriciels que l'on doit entrer dans les conventions du calcul matriciel, auquel cas tous les vecteurs doivent être des "vecteurs colonne". Alors, comme nous l'avons fait ci-dessus, tout "vecteur à plusieurs indices" sera écrit en colonne en utilisant l'ordre lexicographique.

1.3 Régression linéaire multiple

Les deux exemples précédents nous ont montré que l'on pouvait écrire synthétiquement les deux modèles sous une même forme matricielle. Mais cette écriture se généralise également à d'autres modèles. Considérons par exemple l'observation de :

- Y, un vecteur formé de n rendements Y_i d'une réaction chimique (exprimé en pourcentage) ;
- $Z^{(1)}$, un vecteur formé des n mesures $Z_i^{(1)}$ de température de la réaction ;
- $Z^{(2)}$, un vecteur formé des n mesures $Z_i^{(2)}$ de pH du bain de la réaction.

On suppose que le rendement de la réaction chimique (variable à expliquer Y) dépend linéairement des deux variables explicatives, la température et le pH (variables $Z^{(1)}$ et $Z^{(2)}$). On écrit donc le modèle de régression multiple suivant :

$$Y_i = \mu + \beta_1 Z_i^{(1)} + \beta_2 Z_i^{(2)} + \varepsilon_i \tag{3.3}$$

pour $i = 1, \cdots, n$. Il s'en suit l'écriture matricielle suivante :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1^{(1)} & Z_1^{(2)} \\ 1 & Z_2^{(1)} & Z_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & Z_n^{(1)} & Z_n^{(2)} \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

ou encore :

$$Y = X \cdot \theta + \varepsilon \quad \text{avec} \quad \theta = \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{et} \quad X = \begin{pmatrix} 1 & Z_1^{(1)} & Z_1^{(2)} \\ 1 & Z_2^{(1)} & Z_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & Z_n^{(1)} & Z_n^{(2)} \end{pmatrix}. \tag{3.4}$$

2 Le modèle linéaire : définition et hypothèses

Nous avons vu que la régression linéaire simple et l'analyse de la variance à un facteur, qui sont des problèmes relativement différents, conduisaient à des outils statistiques proches : tests de Fisher et de Student, et table d'analyse de la variance. Nous avons vu que matriciellement ces deux modèles s'écrivent de la même manière, ainsi qu'un modèle plus général de régression multiple. Nous allons également montrer que les principaux outils statistiques de ces modèles sont les mêmes.

Définition fondamentale : nous dirons qu'une variable Y constituée de n observations Y_i suit un modèle linéaire statistique si on peut écrire que :

$$Y = X \cdot \theta + \varepsilon \tag{3.5}$$

où

i. X est une matrice constituée de nombres réels, connue, à n lignes et un certain nombre de colonnes que l'on notera k, où $k < n$. Dans le cas du modèle de régression linéaire, chaque colonne représente les données relatives à une variable explicative, sauf pour le cas très fréquent où l'une des colonnes est constituée uniquement de 1 : cela correspondra à la constante (l"intercept") du modèle linéaire. Dans le cas de l'analyse de la variance à un ou plusieurs facteurs (voir le chapitre 5), chaque colonne de la matrice X sera constituée de 1 et de 0, qui correspondront à la présence ou à l'absence de chaque effet. Pour simplifier, on supposera (sauf contre-indication explicite) que **la matrice X est régulière (c'est-à-dire de rang k)**. Ceci implique notamment que pour un vecteur $C \in \mathbb{R}^k$, $X \cdot C = 0 \implies C = 0.$, donc qu'il existe un unique vecteur θ associé au modèle 3.5. Cela implique également que les colonnes de X forment des vecteurs linéairement indépendants de $\mathbb{R}^n$.

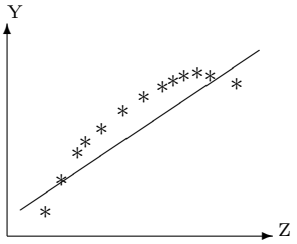
ii. θ est un vecteur inconnu constitué de k réels qui sont des paramètres du modèle ;

iii. le vecteur aléatoire ε appelé erreur du modèle est tel que $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n}$ vérifie les 4 **postulats suivants (hypothèses que l'on accepte et que l'on ne peut tester mais dont la véracité peut être plus ou moins évaluée graphiquement.)** :

- **P1 : les erreurs sont centrées (leurs espérances sont nulles), soit**

$$\mathbb{E}(\varepsilon) = 0.$$

En clair, cela veut dire que le modèle posé (3.5) est correct, que l'on n'a pas oublié un terme pertinent. Considérons le contre-exemple suivant (en régression simple) :



$\Longrightarrow$ il semble qu'un terme quadratique ait été oublié et le modèle devrait plutôt s'écrire : $Y_i = \mu + \beta_1 \cdot Z_i + \beta_2 \cdot (Z_i)^2 + \varepsilon_i$. Si on considère à la place un modèle linéaire simple (comme cela est fait sur la représentation graphique), le vecteur d'erreur ne sera pas centré.

- **P2 : la variance des erreurs est constante (postulat dit d'homoscédasticité), soit :**

$$\mathrm{Var}(\varepsilon_i) = \sigma^2, \quad \text{pour tout } i.$$

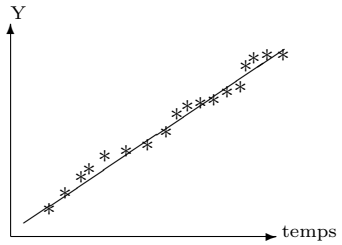
Dans les cas concrets, ce postulat n'est évidemment pas toujours vérifié. Étudions le contre-exemple suivant (issu de l'analyse de variance) : on étudie le taux de suivie des insectes à deux insecticides A et B. On fait différentes répétitions de l'expérience et on obtient le tableau de données suivant :

	taux de survie	
	produit A	produit B
rep1	0,01	0,37
rep2	0.02	0.26
rep3	0.02	0.60
rep4	0.04	0.44
$\vdots$	$\vdots$	$\vdots$

A première vue, l'insecticide A paraît plus efficace que l'insecticide B : le taux de survie à A est plus proche de zéro que celui à B. Mais on remarque surtout que ce taux de survie semble avoir une amplitude de variations beaucoup plus faible avec A qu'avec B : cette dernière constatation, encore appelée présence d'une *hétéroscédasticité*, contredirait l'hypothèse d'homoscédasticité.

- **P3 : les variables ε_i sont indépendantes.**

  On considérera en général que ce postulat est vérifié lorsque chaque donnée correspond à un échantillonnage indépendant ou à une expérience physique menée dans des conditions indépendantes. En revanche, dans des problèmes où le temps joue un rôle important, il est plus difficilement vérifié (une évolution se fait rarement de façon totalement indépendante du passé). Voici un contre-exemple (cas de la régression simple) à ce postulat, visible sur la représentation graphique suivante :



  Dans ce contre-exemple, la variable explicative est le temps et il y a une certaine rémanence ou inertie du phénomène étudié. Cela s'observe par le fait que les données n'oscillent pas en permanence autour de la droite de régression, mais semblent s'attarder une fois qu'elles sont d'un côté.

- **P4 : Les données suivent des lois gaussiennes (ou normales), soit :**

$$\varepsilon_i \sim \mathcal{N}(m_i, \sigma_i^2) \text{ pour tout } i,$$

  où m_i et σ_i^2 sont des paramètres réels. C'est le postulat le moins important, comme nous le verrons plus tard, puisque l'on peut s'en passer quand le nombre de données est important. Remarquons cependant que dans la plupart des cas d'application du modèle linéaire, c'est-à-dire quand les postulats **P1** et **P2** sont vérifiés, l'ajout du postulat **P4** revient à écrire que :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ pour tout } i.$$

  Il n'existe pas vraiment de moyen efficace, graphique ou numérique, pour vérifier si ce postulat de gaussianité est légitime. Pour effectuer un test de gaussianité classique : test de Kolmogorov-Smirnov , Shapiro-Wilks (voir Lecoutre et Tassi [39]), il faudrait pouvoir observer les ε_i ce qui n'est pas le cas. Nous verrons un peu plus loin qu'un histogramme ou une droite de Henri ("QQ-plot") sur les résidus studentisés de la régression peuvent apporter des indications sur cette potentielle normalité des erreurs. Ainsi, notre attitude par rapport à ce

postulat sera ambiguë : par défaut nous le supposerons mais parfois nous ne le supposerons pas, et nous parlerons alors de "modèle linéaire non gaussien". Cependant, comme nous le développerons en détail au chapitre 8, ce postulat n'a plus réellement d'importance lorsque le nombre de données est grand. En pratique, on considérera que cela se produit dès que le nombre de données est de l'ordre de 2 ou 3 dizaines.

3 Formules fondamentales

3.1 Le modèle linéaire en 4 formules

On se place maintenant dans le cadre du modèle linéaire statistique (3.5) auquel, pour commencer, on adjoint les postulats **P1-4**. Voici maintenant les 4 formules fondamentales (notées **F1-4** par la suite) que l'on peut associer à ce modèle. Nous remarquerons que ces formules sont très facilement implémentables et donc calculables par un ordinateur. De plus, elles sont communes à l'analyse de la variance et à la régression. Enfin, elles sont issues de la minimisation en θ de la somme des carrés des résidus (SCR) (ou somme des carrés résiduelle), somme qui peut s'écrire matriciellement sous la forme :

$$\mathrm{SCR}(\theta) = \| Y - X \cdot \theta \|^2 = (Y - X \cdot \theta)' \cdot (Y - X \cdot \theta),$$

où, dans toute la suite, M' désigne la matrice transposée d'une matrice M quelconque. On a alors :

- **F1 :** $\widehat{\theta} = (X' \cdot X)^{-1} \cdot X' \cdot Y$.

  Cette formule fournit l'expression de l'estimateur $\widehat{\theta}$ de θ par moindres carrés (attention, en général θ est un vecteur de paramètres). Un tel calcul est réalisable par un ordinateur qui sait faire du calcul matriciel (on peut traiter sans problème des problèmes comprenant plusieurs centaines de paramètres). Le vecteur Y est gaussien et donc par linéarité $\widehat{\theta}$ l'est également. On peut écrire de manière équivalente que :

  $$X' \cdot X \cdot \widehat{\theta} = X' \cdot Y.$$

  Cette équation vectorielle est un système de k équations scalaires appelées *équations normales.*

- **F2 :** $\mathbb{E}(\widehat{\theta}) = \theta$.

  Cette formule se traduit par le fait que l'estimateur est sans biais. Par des

techniques faisant appel à la notion d'exhaustivité, et sous les postulats **P1-4**, l'estimateur $\widehat{\theta}$ ainsi défini a des propriétés d'optimalité : c'est un estimateur optimal parmi les estimateurs sans biais de θ. En effet, soit $\widetilde{\theta}$ un autre estimateur sans biais. Alors pour toute combinaison linéaire $C' \cdot \theta$, où $C \in \mathbb{R}^k$,

$$\mathrm{Var}\,(C' \cdot \widetilde{\theta}) \geq \mathrm{Var}\,(C' \cdot \widehat{\theta}).$$

- **F3 :** $\mathrm{Var}\,(\widehat{\theta}) = \sigma^2 (X' \cdot X)^{-1}$.

  Cette formule fournit l'expression de la matrice de variance-covariance de l'estimateur $\widehat{\theta}$. Elle permet d'apprécier la précision de cet estimateur.

- **F4 :** $SCR(\widehat{\theta}) = (Y - X \cdot \widehat{\theta})'.(Y - X \cdot \widehat{\theta}) = \| Y - X \cdot \widehat{\theta} \|^2 = \| Y - \widehat{Y} \|^2$ est une variable aléatoire indépendante de $\widehat{\theta}$ et suit une loi $\sigma^2 \cdot \chi^2(n-k)$ (loi du Khi-deux à $n-k$ degrés de liberté, multipliée par le coefficient σ^2, X étant une matrice de taille (n,k)). Cela permet d'estimer σ^2 par l'intermédiaire du carré moyen résiduel

  $$\widehat{\sigma^2} = \mathrm{CMR} = \frac{\mathrm{SCR}(\widehat{\theta})}{n-k} = \frac{\| Y - \widehat{Y} \|^2}{n-k},$$

  qui est proche de σ^2 quand $n-k$ est grand (voir le chapitre 8). De plus $\mathbb{E}\,(\widehat{\sigma^2}) = \sigma^2$, cet estimateur est sans biais. On montre qu'il est optimal comme pour $\widehat{\theta}$.

<u>*Démonstration*</u> : Nous allons utiliser la notion de projection orthogonale pour montrer les différentes formules **F1-4**. Rappelons auparavant que, si u est un vecteur de $\mathbb{R}^n$ et E un sous-espace vectoriel de $\mathbb{R}^n$, alors le projeté orthogonal de u sur E, noté $P_E u$, est le vecteur de E qui minimise $\| u - v \|^2$ pour tout $v \in E$, soit encore

$$\| u - P_E u \|^2 = \min_{v \in E} \| u - v \|^2$$

(en effet, si v est un vecteur quelconque de E, $\| u - v \|^2 = \| u - P_E u \|^2 + \| P_E u - v \|^2 \geq \| u - P_E u \|^2$), d'après le Théorème de Pythagore. Dans toute la suite, on notera $[X]$ le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les vecteurs colonnes constituant la matrice X, ou encore, $[X] = \{X \cdot \theta,\ \theta \in \mathbb{R}^k\}$, pour X une matrice de taille (n,k).

F1 : le vecteur $\widehat{Y} = X \cdot \widehat{\theta} = P_{[X]} Y \in \mathbb{R}^n$, minimisant $\| Y - Y_X \|^2$ pour tout $Y_X \in [X]$, est donc le projeté orthogonal de Y sur $[X]$. Soit maintenant $X^{(i)}$ le $i$ème vecteur colonne de X. Par définition de la projection orthogonale, pour tout $i = 1, \cdots, k$,

$$< X^{(i)}, Y > = (X^{(i)})' \cdot Y = < X^{(i)}, X \cdot \widehat{\theta} >,$$

où $< ., . >$ note le produit scalaire euclidien usuel dans $\mathbb{R}^n$. En mettant les équations les unes au-dessus des autres, pour tout $i = 1, \cdots, k$, on obtient

$$X' \cdot Y = (X' \cdot X) \cdot \widehat{\theta} \implies \widehat{\theta} = (X' \cdot X)^{-1} \cdot X' \cdot Y,$$

car X est supposée être une matrice régulière (donc la matrice $X' \cdot X$ est inversible).

F2 : en utilisant la linéarité de l'espérance, on a

$$\mathbb{E}(\widehat{\theta}) = \mathrm{E}\left[(X' \cdot X)^{-1} \cdot X' \cdot Y\right] = (X' \cdot X)^{-1} X' \cdot \mathbb{E}(Y) = (X' \cdot X)^{-1} \cdot X' \cdot (X \cdot \theta) = \theta.$$

On montre par ailleurs que le modèle linéaire, au prix d'un changement de paramétrisation, est un modèle exponentiel dont la statistique naturelle est exhaustive minimale et complète. Le théorème de Rao-Blackwell implique que $\widehat{\theta}$ est optimal comme estimateur sans biais fonction de cette statistique exhaustive minimale et complète. Nous renvoyons pour les détails à des ouvrages plus théoriques comme Dacunha-Castelle et Duflo [20] p. 174-175 ou Coursol [19] p. 13-14.

F3 : comme $Y = X \cdot \theta + \varepsilon$, d'après les postulats sur ε, $\mathrm{Var}(Y) = \sigma^2 \cdot I_n$, où I_n désigne la matrice identité sur $\mathbb{R}^n$. Ainsi,

$$\mathrm{Var}(\widehat{\theta}) = (X' \cdot X)^{-1} \cdot X' \cdot (\mathrm{Var}(Y)) \cdot X \cdot (X' \cdot X)^{-1} = \sigma^2 \cdot (X' \cdot X)^{-1}.$$

F4 : par linéarité de la projection orthogonale, $P_{[X]}Y = X \cdot \theta + P_{[X]}\varepsilon$. Ainsi,

$$\mathrm{SCR}(\widehat{\theta}) = \| Y - P_{[X]}Y \|^2 = \| \varepsilon - P_{[X]}\varepsilon \|^2 .$$

Or, on peut encore écrire que $\varepsilon - P_{[X]}\varepsilon = P_{[X]^\perp}\varepsilon$, où $[X]^\perp$ désigne le sous-espace vectoriel de $\mathbb{R}^n$ orthogonal de $[X]$. La dimension de $[X]^\perp$ étant $n - k$, d'après les résultats sur les variables gaussiennes indépendantes (Théorème de Cochran dont un rappel est proposé en annexe), la variable aléatoire $SCR(\widehat{\theta})$ a pour distribution une loi $\sigma^2 \cdot \chi^2(n-k)$. Une variable aléatoire distribuée suivant la loi $\chi^2(n-k)$ pouvant encore s'écrire comme la somme de $(n-k)$ carrés de variables gaussiennes centrées réduites indépendantes, il est clair que $\mathbb{E}(\widehat{\sigma^2}) = \sigma^2$. Par les mêmes arguments d'exhaustivité que précédemment, la propriété de non-biais implique l'optimalité.

Enfin, comme $X \cdot \widehat{\theta} = X \cdot \theta + P_{[X]}\varepsilon$, toujours en utilisant le Théorème de Cochran, comme $P_{[X]}\varepsilon$ et $P_{[X]^\perp}\varepsilon$ sont des projections sur des sous-espaces orthogonaux de $\mathbb{R}^n$, alors $P_{[X]}\varepsilon$ est indépendant de $SCR(\widehat{\theta})$, donc $\widehat{\theta}$ et $\widehat{\sigma^2}$ sont également indépendants. ■

Remarque : En absence du postulat **P4**, l'estimateur $\widehat{\theta}$ reste optimal parmi les estimateurs linéaires sans biais (voir Exercices).

3.2 Un exemple : les équations explicites dans le cas de la régression linéaire simple.

On considère le cas de la régression linéaire simple, cas particulier du modèle (1.1), auquel on adjoint les postulats **P1-4**. Nous allons retrouver tous les résultats du chapitre 1 à partir de leurs écritures matricielles présentées précédemment.

Soit $\overline{Z}$ la moyenne des Z et on pose $Z_i^0 = Z_i - \overline{Z}$ (régresseur centré). On écrit alors le modèle :

$$Y_i = \mu + \beta \cdot \overline{Z} + \beta \cdot (Z_i - \overline{Z}) + \varepsilon_i,$$

donc, en posant $\mu' = \mu + \beta \cdot \overline{Z}$, on obtient,

$$Y_i = \mu' + \beta \cdot Z_i^o + \varepsilon_i.$$

Nous allons voir que travailler avec les Z_i^o au lieu des Z_i permet de se placer dans le cadre de l'orthogonalité et de simplifier les calculs (nous reviendrons plus en détail sur cette notion d'orthogonalité dans le chapitre 7). Le modèle s'écrit alors matriciellement sous la forme

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1^o \\ \vdots & \vdots \\ 1 & Z_n^o \end{pmatrix} \begin{pmatrix} \mu' \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Ainsi, en posant $X = \begin{pmatrix} 1 & Z_1^o \\ \vdots & \vdots \\ 1 & Z_n^o \end{pmatrix}$ et $\theta = \begin{pmatrix} \mu' \\ \beta \end{pmatrix}$, on peut utiliser les formules **F1-4** précédentes. Or,

$$X' \cdot X = \begin{pmatrix} n & \sum Z_i^o \\ \sum Z_i^o & \sum (Z_i^o)^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & \sum (Z_i^o)^2 \end{pmatrix} \text{ d'où } (X' \cdot X)^{-1} = \begin{pmatrix} n^{-1} & 0 \\ 0 & \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}.$$

De plus,

$$X' \cdot Y = \begin{pmatrix} 1 & \cdots & 1 \\ Z_1^o & \cdots & Z_n^o \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i Z_i^o \end{pmatrix}$$

et donc

$$\widehat{\theta} = \begin{pmatrix} \widehat{\mu'} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} n^{-1} & 0 \\ 0 & \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix} \begin{pmatrix} \sum Y_i \\ \sum Y_i Z_i^o \end{pmatrix} = \begin{pmatrix} n^{-1} \sum Y_i \\ \left(\sum Y_i Z_i^o\right) \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}.$$

On a donc également :

$$\text{Var} \begin{pmatrix} \widehat{\mu'} \\ \widehat{\beta} \end{pmatrix} = \sigma^2 (X' \cdot X)^{-1} = \begin{pmatrix} \sigma^2 n^{-1} & 0 \\ 0 & \sigma^2 \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}$$

Or, $\widehat{\mu} = \widehat{\mu'} - \widehat{\beta} \cdot \bar{Z}$, et donc :

$$\mathrm{Var}\begin{pmatrix} \widehat{\mu} \\ \widehat{\beta} \end{pmatrix} = \begin{pmatrix} \mathrm{Var}\,(\widehat{\mu'}) + (\overline{Z})^2 \mathrm{Var}\,(\widehat{\beta}) - 2\overline{Z}\mathrm{cov}\,(\widehat{\mu'}, \widehat{\beta}) & \mathrm{cov}\,((\widehat{\mu'} - \widehat{\beta}\overline{Z}), \widehat{\beta}) \\ \mathrm{cov}\,((\widehat{\mu'} - \widehat{\beta}\overline{Z}), \widehat{\beta}) & \sigma^2 \left(\sum (Z_i^o)^2\right)^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 n^{-1} + \sigma^2 \cdot \overline{Z}^2 \left(\sum (Z_i - \overline{Z})^2\right)^{-1} & -\overline{Z} \cdot \sigma^2 \left(\sum (Z_i - \overline{Z})^2\right)^{-1} \\ -\overline{Z} \cdot \sigma^2 \left(\sum (Z_i - \overline{Z})^2\right)^{-1} & \sigma^2 \left(\sum (Z_i - \overline{Z})^2\right)^{-1} \end{pmatrix}$$

Nous avons donc retrouvé toutes les formules que nous avions admises au chapitre 1.

4 Tests fondamentaux et intervalles de confiance

4.1 Tests de Fisher d'un sous-modèle

Les tests de Fisher sont communs à l'analyse de la variance et à la régression. Au cours du chapitre précédent, nous avons vu deux exemples de sous-modèles :

Modèle général de la régression linéaire simple :	$Y_i = \mu + \beta \cdot Z_i + \varepsilon_i$.
Sous-modèle avec nullité de la pente :	$Y_i = \mu + \varepsilon_i$.
Modèle général de l'analyse de la variance à 1 facteur :	$Y_{ij} = \mu_i + \varepsilon_{ij}$.
Sous-modèle avec égalité des groupes :	$Y_{ij} = \mu + \varepsilon_{ij}$.

Plaçons nous maintenant dans le cadre général du modèle linéaire. Soit le modèle (3.5), où X est une matrice de rang $k < n$. On note SCR la somme des carrés résiduelle de ce modèle, associée à $n - k$ degrés de liberté, soit

$$\mathrm{SCR} = \| \, Y - X \cdot \widehat{\theta} \, \|^2 .$$

On considère le sous-modèle issu du modèle (3.5) et tel que

$$Y = X^{(0)} \cdot \theta^{(0)} + \varepsilon, \tag{3.6}$$

où $[X^{(0)}] \subset [X]$, $\dim[X^{(0)}] = k_0 < k = dim[X]$. Le plus souvent, $X^{(0)}$ est la matrice constituée de k_0 vecteurs colonnes de X (avec $k_0 < k$) et $\theta^{(0)}$ est un vecteur de longueur k_0. On note alors SCR_0 la somme des carrés résiduelle de ce sous-modèle, associée donc à $(n - k_0)$ degrés de liberté, soit

$$\mathrm{SCR}_0 = \| \, Y - X^{(0)} \cdot \widehat{\theta^{(0)}} \, \|^2 .$$

On pose

$$Y = R + \varepsilon$$

et on veut tester l'hypothèse nulle définie par le modèle (3.6) contre le modèle général. C'est-à-dire

$$\text{contre} \quad \begin{array}{lcl} H_0 & : & R \in [X^{(0)}] \\ H_1 & : & R \in [X] \setminus [X^{(0)}] \end{array} .$$

Proposition 3.1 *Dans le cadre du modèle linéaire général (3.5) auquel on a adjoint les postulats* **P1-4**, *et avec les notations précédentes, sous l'hypothèse nulle H_0 (le sous-modèle (3.6) est vrai), alors :*

$$\widehat{F} = \frac{(SCR_0 - SCR)/(k - k_0)}{SCR/(n-k)}$$

suit une loi de Fisher de paramètres $(k - k_0, n - k)$. De plus, $\widehat{F}$ est indépendante de $\widehat{Y}^{(0)} = X^{(0)} \cdot \widehat{\theta}^{(0)}$ (calculé sous l'hypothèse H_0).

Démonstration : Pour aider à la compréhension de la démonstration, on pourra se reporter à la figure 4.1. On se place sous l'hypothèse nulle. On a

$$\text{SCR} = \|Y - P_{[X]}Y\|^2 = \|P_{[X]^\perp}Y\|^2 = \|P_{[X]^\perp}\varepsilon\|^2 = \|V\|^2,$$

car $P_{[X]^\perp}(R) = 0$ (car $R \in [X]$), $V = P_{[X]^\perp}\varepsilon$ étant indiqué sur la figure 4.1. De même,

$$\text{SCR}_0 = \|Y - P_{[X^{(0)}]}Y\|^2 = \|P_{[X^{(0)}]^\perp}Y\|^2 = \|P_{[(X^{(0)})]^\perp}\varepsilon\|^2 = \|U\|^2,$$

avec $U := P_{[(X^{(0)})]^\perp}\varepsilon$. Comme $[X^{(0)}] \subset [X]$, on a $[X]^\perp \subset [X^{(0)}]^\perp$. Soit maintenant A le sous-espace vectoriel de $[X]$, complémentaire et orthogonal à $[X^{(0)}]$. Cet espace est de dimension $k - k_0$. On a

$$A \overset{\perp}{\oplus} [X^0] = [X].$$

Posons $W = P_A\varepsilon$ (voir la figure 4.1, où est également indiqué $U0 = P_{[X^{(0)}]}\varepsilon$). Par le Théorème de Pythagore, on a $\|U\|^2 = \|V\|^2 + \|W\|^2$, ou encore

$$\|P_{[X^{(0)}]^\perp}\varepsilon\|^2 = \|P_{[X]^\perp}\varepsilon\|^2 + \|P_A\varepsilon\|^2.$$

Comme ε est un vecteur composé de variables gaussiennes centrées, de variance σ^2 indépendantes, d'après le Théorème de Cochran, ses projections sur deux sous-espaces

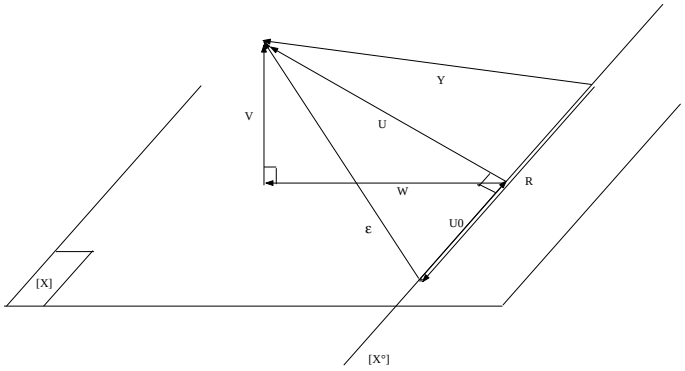


FIGURE 3.1 – Vecteurs intervenant dans le test de Fisher

orthogonaux sont indépendantes et leurs normes ont pour distributions des lois $\sigma^2 \cdot \chi^2$. On en déduit ainsi que

$$\begin{aligned} \text{SCR} &= \|P_{[X]^\perp}\varepsilon\|^2 \text{ suit la loi } \sigma^2 \cdot \chi^2(n-k), \\ \text{SCR}_0 - \text{SCR} &= \|P_{[X^{(0)}]^\perp}\varepsilon\|^2 - \|P_{[X]^\perp}\varepsilon\|^2 = \|P_A\varepsilon\|^2 \text{ suit la loi } \sigma^2 \cdot \chi^2(k-k_0). \end{aligned}$$

et ces deux quantités sont indépendantes. Donc

$$\widehat{F} = \frac{\|P_A\varepsilon\|^2/(k-k_0)}{\|P_{[X]^\perp}\varepsilon\|^2/(n-k)}$$

suit bien la loi de Fisher de paramètre $(k-k_0, n-k)$.

De plus, comme le sous-espace $[X^{(0)}]$ est orthogonal à $[X]^\perp$ et à A, on en déduit que $\widehat{Y}^{(0)} = P_{[X^{(0)}]}Y = X^{(0)} \cdot \theta^{(0)} + P_{[X^{(0)}]}\varepsilon = X^{(0)} \cdot \widehat{\theta}^{(0)}$ est indépendant de SCR et de $\text{SCR}_0 - \text{SCR}$, donc on en déduit que $\widehat{F}$ est indépendant de $\widehat{Y}^{(0)}$ et de $\widehat{\theta}^{(0)}$. ■

Remarque : Dans le cas particulier de la régression linéaire multiple, pour mesurer l'adéquation globale des données au modèle, on utilise souvent *le coefficient de détermination* R^2 défini par (1.4) (même définition que pour la régression linéaire simple, voir chapitre 1), soit

$$R^2 = \frac{\sum_{i=1}^n (\overline{Y}_i - \widehat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \overline{Y}_i)^2} = \frac{\|\overline{Y} - \widehat{Y}\|}{\|Y - \overline{Y}\|}.$$

En reprenant le fil de la démonstration précédente, on comprend bien que le numérateur et le dénominateur de ce coefficient suivent des lois du Khi-deux décentrées non

indépendantes. Le coefficient R^2 ne suit donc pas une des lois usuelles simples. De plus, ce coefficient est quelque peu trompeur. En effet, plus nombreux sont les régresseurs utilisés, plus il est proche de 1, et donc plus forte semble l'adéquation du modèle aux données. Pourtant rien ne garantit la légitimité de la présence de chacun des régresseurs. Un test de Fisher sera alors beaucoup plus probant, en attendant de voir les méthodes du chapitre 9.

4.2 Test de Student de la nullité d'une combinaison linéaire

On se place à nouveau dans le cadre du modèle linéaire standard (3.5). Soit une combinaison linéaire $C' \cdot \theta$ du vecteur de paramètres θ, avec $C \in \mathbb{R}^k$ (à titre d'exemple, une telle combinaison linéaire pourrait être $\mu_1 - \mu_2$ en analyse de la variance, ou β en régression linéaire simple). On veut tester la nullité de $C' \cdot \theta$, donc tester :

$$\text{contre} \quad \begin{array}{rcl} H_0 & : & C' \cdot \theta = 0 \\ H_1 & : & C' \cdot \theta \neq 0 \end{array} .$$

Proposition 3.2 *Dans le cadre du modèle linéaire général (3.5) auquel on a adjoint les postulats* **P1-4**, *et avec les notations précédentes, sous l'hypothèse nulle* H_0 *(*$C' \cdot \theta = 0$*), alors :*

$$\widehat{T} = \frac{C' \cdot \widehat{\theta}}{\sqrt{\widehat{\sigma}^2 \cdot C' \cdot (X' \cdot X)^{-1} C}}$$

suit une loi de Student de paramètre $(n - k)$.

Démonstration : Dans le cadre du modèle linéaire général (3.5) et de ses différents postulats, alors d'après **F2** et **F3**,

$$\text{Var}\,(C' \cdot \widehat{\theta}) = \sigma^2 \cdot C' \cdot (X' \cdot X)^{-1} \cdot C.$$

Vu ce qui précède, on estime assez naturellement $\text{Var}\,(C' \cdot \widehat{\theta})$ par $\widehat{\sigma^2} \cdot C' \cdot (X' \cdot X)^{-1} \cdot C$. On divise alors $C' \cdot \widehat{\theta}$ par son écart-type estimé. Une telle démarche est classique en statistique ; on dit encore que l'on a studentisé la statistique. On obtient ainsi la statistique $\widehat{T}$ qui s'écrit encore,

$$\widehat{T} = \frac{C' \cdot \widehat{\theta}}{\sqrt{\sigma^2 \cdot C' \cdot (X' \cdot X)^{-1} \cdot C}} \times \frac{\sqrt{\sigma^2}}{\sqrt{\widehat{\sigma}^2}}.$$

Or, sous l'hypothèse H_0, comme $C' \cdot \theta = 0$, $C' \cdot \widehat{\theta}$ est une variable gaussienne centrée, et il est clair que

$$\frac{C' \cdot \widehat{\theta}}{\sqrt{\sigma^2 \cdot C' \cdot (X' \cdot X)^{-1} \cdot C}} \overset{\mathcal{L}}{\sim} \mathcal{N}(0, 1).$$

De plus, $\frac{n-k}{\sigma^2}\widehat{\sigma}^2 = \frac{n-k}{\sigma^2}\|P_{[X]^\perp}\varepsilon\|^2$ suit une loi $\chi^2(n-k)$ et est indépendant de $C'\cdot P_{[X]}\varepsilon$ puisque les deux sous-espaces sont orthogonaux. Par définition d'une loi de Student, on en déduit que $\widehat{T}$ suit bien une loi de Student à $(n-k)$ degrés de liberté. ■

Cette proposition permet de réaliser un test de Student des hypothèses H_0 et H_1 précédentes. En effet, si $|\widehat{T}| > T_{n-k,1-\alpha/2}$, où $T_{n-k,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $(n-k)$ degrés de liberté, alors on rejette l'hypothèse H_0 et donc la nullité de $C'\cdot\theta$. Dans le cas contraire, on acceptera H_0.

Remarque : $C'\cdot\theta = 0$ définit un sous-modèle (pas toujours facile à écrire) du modèle linéaire général(3.5) et le test de Fisher associé est exactement le même que le test de Student ci-dessus. Il s'agit simplement d'une présentation différente.

4.3 Test de Fisher de la nullité jointe de plusieurs combinaisons linéaires

On suppose que l'on réalise une expérience de type médicale, avec un seul facteur "traitement" possédant 5 niveaux. On peut, par exemple, écrire le modèle sous la forme $Y_{ij} = \theta_i + \varepsilon_{ij}$ pour $i = 1,\cdots,5$, ou bien sous la forme générale (3.5) avec $\theta = (\theta_i)_{1\leq i\leq 5}$.

Supposons que l'on veuille tester l'hypothèse $H_0 : \theta_1 = \theta_2 = \theta_3$ et $\theta_4 = \theta_5$, contre l'hypothèse H_1.
Notons

$$C' = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \quad \text{alors} \quad C'\cdot\theta = \begin{pmatrix} \theta_1-\theta_2 \\ \theta_2-\theta_3 \\ \theta_4-\theta_5 \end{pmatrix}.$$

Ainsi tester H_0 revient à tester l'hypothèse nulle $C'\cdot\theta = 0$.

Plus généralement, on se place dans le cadre du modèle linéaire (3.5) et ses postulats, et on considère C une matrice réelle connue de taille (k,p), que l'on supposera de rang $p \leq k$. On désire tester :

$$\text{contre} \quad \begin{array}{ll} H_0 & : \quad C'\cdot\theta = 0 \\ H_1 & : \quad C'\cdot\theta \neq 0 \end{array}.$$

Proposition 3.3 *Dans le cadre du modèle linéaire général (3.5) auquel on a adjoint*

les postulats **P1-4**, *et avec les notations précédentes, sous l'hypothèse nulle* H_0, *alors :*

$$\widehat{F} = \frac{\widehat{\theta}' \cdot C \cdot (C' \cdot (X' \cdot X)^{-1} \cdot C)^{-1} C' \cdot \widehat{\theta}}{p \cdot \widehat{\sigma}^2}$$

suit une loi de Fisher de paramètre $(p, n-k)$.

Démonstration : Sous l'hypothèse H_0 et comme dans la démonstration précédente, $C' \cdot \widehat{\theta} = C' \cdot P_{[X]} \cdot \varepsilon$ est un vecteur gaussien centré de matrice de variance-covariance $V = \sigma^2 \cdot C' \cdot (X' \cdot X)^{-1} \cdot C$. Or la matrice V ainsi définie est symétrique et inversible, car C est supposée être de rang p. Comme V est diagonalisable avec des valeurs propres positives, il est donc évident qu'il existe une "racine carrée de son inverse" : il existe $V^{-1/2}$ symétrique telle que

$$V^{-1/2} \cdot V^{-1/2} \cdot V = V^{-1/2} \cdot V \cdot V^{-1/2} = I_d.$$

En conséquence, $V^{-1/2} \cdot C' \cdot \widehat{\theta}$ est un vecteur gaussien centré de taille p et de matrice de variance-covariance identité. On en déduit alors que

$$\|V^{-1/2} \cdot C' \cdot \widehat{\theta}\|^2 = \widehat{\theta}' \cdot C \cdot V^{-1/2} \cdot V^{-1/2} \cdot C' \cdot \widehat{\theta} = \widehat{\theta}' \cdot C \cdot (\sigma^2 \cdot C' \cdot (Z' \cdot Z)^{-1} \cdot C)^{-1} \cdot C' \cdot \widehat{\theta} \text{ suit une loi } \chi^2(p).$$

De plus, on l'a déjà vu, $(n-k) \cdot \widehat{\sigma^2} = P_{[X]^\perp} \varepsilon$ suit une loi $\sigma^2 \chi^2(n-k)$, en étant indépendant de $C' \cdot \widehat{\theta}$ puisque $[X]$ et $[X]^\perp$ sont orthogonaux. On en déduit alors que $\dfrac{\|V^{-1/2} \cdot C' \cdot \widehat{\theta}\|^2}{p} \times \dfrac{\sigma^2}{\widehat{\sigma}^2} = \widehat{F}$ suit une loi de Fisher de paramètre $(p, n-k)$. ■

Comme précédemment, on en déduit un test (dit test de Fisher ou test F) sur les hypothèses H_0 et H_1 : on rejettera notamment H_0 lorsque $\widehat{F} > F_{(p,n-k),1-\alpha}$.

Remarque : Ce test est une généralisation assez naturelle du test précédent au cas où plusieurs combinaisons linéaires sont nulles conjointement. On montre (voir par exemple Coursol, 1980) que ce test est exactement le même que le test de Fisher défini par des différences de somme des carrés résiduelles avec le sous-modèle linéaire défini par l'hypothèse $C' \cdot \theta = 0$.

4.4 Intervalles et régions de confiance

Commençons par les intervalles de confiance pour une combinaison linéaire $C' \cdot \theta$. On reprend les notations de la partie 4.2. Le test de cette partie se généralise en un test de :

$$\text{contre} \quad \begin{array}{rcl} H_0 & : & C' \cdot \theta = c_0 \\ H_1 & : & C' \cdot \theta \neq c_0 \end{array} .$$

qui est maintenant basé sur la statistique :

$$\widehat{T} = \frac{C' \cdot \widehat{\theta} - c_0}{\sqrt{\widehat{\sigma}^2 \cdot C' \cdot (X' \cdot X)^{-1} C}}$$

(la suite étant identique). En statistique, il existe un principe général de correspondance entre les famille de tests et les régions de confiance, voir par exemple le livre P. Toulouse [59]. Une conséquence de ce principe est que l'ensemble des c_0 acceptés par un test de niveau α définit ici un intervalle de confiance au niveau de confiance $1-\alpha$. Il est facile de vérifier que cet intervalle de confiance vaut :

$$IC = \left[C' \cdot \widehat{\theta} - T_{n-k,1-\alpha/2} \sqrt{\widehat{\sigma}^2 \cdot C' \cdot (X' \cdot X)^{-1} C}\, ,\, C' \cdot \widehat{\theta} + T_{n-k,1-\alpha/2} \sqrt{\widehat{\sigma}^2 \cdot C' \cdot (X' \cdot X)^{-1} C} \right].$$

De la même manière, si maintenant comme dans la partie 4.3 qui précède, $C' \cdot \theta$ est de dimension $p > 1$ et si c_0 est une valeur particulière appartenant à $\mathbb{R}^p$, on peut généraliser le test de cette section en un test des hypothèses :

$$\text{contre} \quad \begin{array}{rcl} H_0 & : & C' \cdot \theta = c_0 \\ H_1 & : & C' \cdot \theta \neq c_0 \end{array} .$$

qui est basé maintenant sur la statistique :

$$\widehat{F} = \frac{(\widehat{\theta}' \cdot C - c_0') \cdot (C' \cdot (X' \cdot X)^{-1} \cdot C)^{-1} (C' \cdot \widehat{\theta} - c_0)}{p \cdot \widehat{\sigma}^2}$$

qui suit encore, sous l'hypothèse nulle, une loi de Fisher de paramètres $(p, n-k)$. L'ensemble des c_0 acceptés par ce test au niveau α est maintenant une région de confiance, plus précisément un ellipsoïde RC défini par :

$$RC = \{c \in \mathbb{R}^p : (\widehat{\theta}' \cdot C - c_0) \cdot (C' \cdot (X' \cdot X)^{-1} \cdot C)^{-1} C' \cdot (\widehat{\theta} - c_0) \leq p \cdot \widehat{\sigma}^2 F_{(p,n-k),1-\alpha}\}.$$

Remarque : la méthode de Scheffé qui sera présentée au Chapitre 5, Section 3, est fondée sur les projections dans différentes directions de cet ellipsoïde.

5 Quand les postulats ne sont pas respectés...

Le postulat de gaussianité des erreurs est le plus difficile à vérifier en pratique. Les tests classiques de normalité (test de Kolmogorov-Smirnov, Cramer-Von Mises, Anderson-Darling ou de Shapiro-Wilks) demanderaient l'observation des erreurs ε_i elles-mêmes ; ils perdent beaucoup de puissance quand ils sont appliqués sur les résidus

$\widehat{\varepsilon_i} = (Y - \widehat{Y})_i$, notamment en raison du fait que ces résidus ne sont pas indépendants (en général). On peut cependant toujours faire des droites de Henri ou QQ-plots (voir exemples informatiques) pour mettre en évidence des écarts évidents. Il n'en reste pas moins que le postulat de gaussianité sera le plus souvent un credo que l'on ne pourra pas vraiment vérifier expérimentalement. Fort heureusement, comme nous allons le décrire, il existe une théorie asymptotique (donc pour les grands échantillons) du modèle linéaire qui n'a pas besoin de cette hypothèse. Comme il est dit dans l'introduction, c'est dans cette optique là qu'il faut réellement penser le modèle linéaire. Cette théorie est exposée en détail dans le chapitre 8, nous nous contenterons ici d'en résumer les résultats principaux.

Nous allons d'abord énoncer les propriétés qui restent vraies en absence du postulat **P4** et éventuellement des autres postulats. Donc par la suite, on se place juste dans le cadre du modèle linéaire général (3.5), sans les postulats **P1-4**.

Propriétés de l'estimateur des moindres carrés $\widehat{\theta}$

Nous considérons donc

$$\widehat{\theta} = (X' \cdot X)^{-1} \cdot X' \cdot Y.$$

Nous donnons les résultats sans les démonstrations qui sont le plus souvent faciles et que nous laissons au lecteur à titre d'exercices.

- $\widehat{\theta}$ reste sans biais, $\mathbb{E}(\widehat{\theta}) = \theta$, sous le seul postulat **P1** ;
- la matrice de variance-covariance de $\widehat{\theta}$ reste égale à $\sigma^2(X' \cdot X)^{-1}$ sous les seuls postulats **P2** et **P3**, mais si **P1** n'est pas vrai cette propriété a peu d'intérêt ;
- $\widehat{\theta}$ n'est plus un estimateur optimal parmi les estimateurs sans biais, mais il le reste parmi les estimateurs linéaires sans biais sous **P1-3** ;
- $\widehat{\theta}$ est gaussien sous **P3** et **P4**. Si **P4** n'est pas vrai, il tend a être gaussien pour de grands échantillons (on dit qu'il est asymptotiquement gaussien). Voir le chapitre 8 pour le détail des conditions.

Propriétés de l'estimateur de σ^2

Cette étude n'a bien-sûr d'intérêt que si σ^2 est bien définie ce qui nécessite que le postulat **P2** soit vrai. Nous considérons

$$\widehat{\sigma^2} = \frac{1}{n-k}\|Y - X \cdot \widehat{\theta}\|^2.$$

Alors :

- $\widehat{\sigma^2}$ sous les postulats **P1-3** reste un estimateur sans biais même si le postulat de gaussianité **P4** n'est pas vérifié (voir exercice 6), soit : $\mathbb{E}(\widehat{\sigma^2}) = \sigma^2$;
- il est bien clair en revanche que $\widehat{\sigma^2}$ ne suit plus une loi $\sigma^2/(n-k) \cdot \chi^2(n-k)$ dès que le postulat **P4** n'est pas vérifié ;
- on montre facilement que sous les postulats **P1-3**, $\widehat{\sigma^2}$ converge en probabilité vers σ^2 même si **P4** n'est pas respecté (voir le chapitre 8) ;
- enfin, sous les seuls postulats **P1-3**, dès que la loi de ε admet un moment d'ordre 4, $\widehat{\sigma^2}$ converge à la vitesse $\sqrt{n}$ vers σ^2 (voir le chapitre 8) mais sa vitesse exacte de convergence dépend du type de loi, plus précisément du coefficient de Kurtosis.

Propriétés des statistiques de test F et T

Dans cette partie, on considère simplement le cas du modèle linéaire non-gaussien : sans **P4**. Le résultat général est la validité asymptotique (pour de grands effectifs) des tests évoqués (sous certaines conditions peu restrictives). On trouvera également dans le chapitre 8 les détails et explications théoriques d'un tel résultat.

Pour illustrer cette validité dans un cas simple, prenons l'exemple de l'analyse de la variance à un facteur. Nous avons vu que l'estimateur $\widehat{\mu}_i$ de la valeur d'une classe était une simple moyenne : $\widehat{\mu}_i = Y_{i.}$. Pour "mesurer" la vitesse de convergence de $\widehat{\mu}_i$ vers μ_i, on utilise le Théorème de la Limite Centrale. Pour revenir au problème précédent d'analyse de la variance à un facteur, une conséquence de ce théorème est que pour n grand, $\widehat{\mu}_i = Y_{i.}$ suit approximativement une loi gaussienne et une conséquence de la loi des grands nombres est que $\widehat{\sigma^2}$ est très proche de σ^2. Ceci implique par exemple que si l'on veut tester l'hypothèse nulle "$\mu_i = m$", pour un i donné et avec m un réel connu, alors comme précédemment on considérera le test de Student dont la statistique est :

$$\widehat{T} = \frac{\widehat{\mu}_i - m}{\sqrt{\widehat{\sigma}^2/n}}.$$

Pour n grand, $\widehat{T}$ suit approximativement une loi gaussienne centrée réduite, qui n'est autre que la limite d'une loi de Student $T(n)$ dont le nombre n de degrés de liberté tend vers l'infini (voir également le chapitre 8).

Ce résultat théorique peut être complété par une étude par simulation. Dans un

mémoire, Bonnet et Lansiaux [15] ont étudié le comportement du test de Fisher en analyse de la variance à un facteur à 2, 5 ou 10 niveaux, avec des indices de répétition de 2, 4 ou 8 ; on a donc de 4 à 80 données dans chaque expérience. La validité du test est appréciée par le niveau réel du test pour un niveau nominal de 10 %, 5 % ou 1 %. Divers types de loi non normales sont utilisées. On aperçoit un écart au comportement nominal du test seulement dans le cas où tous les éléments suivants sont réunis :

- dispositifs déséquilibrés,
- petits échantillons,
- loi dissymétrique.

Dans les autres cas tout se passe comme si les données étaient gaussiennes. Ainsi cette étude confirme, que sauf cas extrêmes, le test de Fisher n'a pas besoin de l'hypothèse de gaussianité pour être approximativement exact.

Modèles avec corrélations

Il est possible de modéliser des corrélations entre erreurs, par exemple en supposant que ces erreurs sont issues d'un processus ARMA, ce qui permet de ne plus avoir besoin du postulat **P3**. Nous ne détaillerons pas cette partie plus complexe et renvoyons à des ouvrages d'économétrie dans lesquelles ceci est souvent traité (par exemple, Amemiya [3], Green [30], Guyon [31] ou Jobson [35]). Il est possible également de modéliser les liaisons par des modèles a effets aléatoires et poser un modèle mixte. Ceci sera fait dans le chapitre 10.

6 Exercices

Exercice 3.1

(*) Soit Y qui suit un modèle linéaire non-gaussien, et soit $T \in \mathbb{R}^n$ un vecteur déterministe. Montrer que

$$\mathbb{E}\left(\|T - Y\|^2\right) = n\sigma^2 + \|T - X\theta\|^2.$$

Exercice 3.2

(**) [Théorème de Gauss-Markov] On se propose de montrer dans cet exercice que, pour le modèle linéaire non gaussien (avec les postulats **P1-3** mais sans **P4**), l'estimateur des moindres carrés $\widehat{\theta}$ reste optimal mais maintenant seulement parmi les

estimateurs linéaires sans biais.

L'optimalité veut dire que si $\widetilde{\theta}$ est un autre estimateur linéaire sans biais :

$$\mathrm{Var}\,(\widetilde{\theta}) - \mathrm{Var}\,(\widehat{\theta}) \text{ est une matrice semi-définie positive,}$$

ou encore, ce qui est équivalent, que pour toute combinaison linéaire $C' \cdot \theta$ des paramètres,

$$\mathrm{Var}\,(C' \cdot \widetilde{\theta}) \geq \mathrm{Var}\,(C' \cdot \widehat{\theta}).$$

i. Posons $\widetilde{\theta} = M \cdot Y$ où M est une matrice de taille (k, n). Montrer que $M \cdot X = I_n$.

ii. Ecrire $\widehat{\theta} = T \cdot P_{[X]}Y$, et montrer que $M \cdot P_{[X]} = T \cdot P_{[X]}$.

iii. Montrer que $\widetilde{\theta} = \widehat{\theta} + M \cdot P_{[X]^\perp}Y$, la somme étant non-corrélée. Conclure.

Exercice 3.3

(**) Soit θ_1 et θ_2 deux paramètres réels inconnus et soit :

- Y_1 un estimateur sans biais de $\theta_1 + \theta_2$ et de variance σ^2 ;
- Y_2 un estimateur sans biais de $2\theta_1 - \theta_2$ et variance $4\sigma^2$;
- Y_3 un estimateur sans biais de $6\theta_1 + 3\theta_2$ et de variance $9\sigma^2$,

les estimateurs Y_1, Y_2 et Y_3 étant indépendants. Quels estimateurs de θ_1 et θ_2 proposeriez-vous ? (on pourra utiliser l'exercice précédent).

Exercice 3.4

(**) [Estimation de la variance] On considère un modèle linéaire non-gaussien (sans **P4**). On veut déterminer l'espérance de $\widehat{\sigma^2}$.

i. Montrer que $(n-k)\widehat{\sigma^2} = \mathrm{Tr}(\varepsilon' \cdot P_{[X]\perp}\varepsilon)$ où Tr désigne la trace.

ii. En utilisant le fait que $\mathrm{Tr}(A \cdot B) = \mathrm{Tr}(B \cdot A)$, montrer que $(n-k)\mathbb{E}\,(\widehat{\sigma^2}) = \sigma^2 Tr(P_{[X]\perp}\varepsilon' \cdot \varepsilon)$.

iii. Conclure.

Exercice 3.5

(*) [Maximum de vraisemblance] Reprendre l'exercice 5 du chapitre 1 dans le cadre général du modèle linéaire (3.5).

Exercice 3.6

(*) [Test du rapport de vraisemblance] Reprendre l'exercice 5 du chapitre 1 dans le cadre général du modèle linéaire (3.5).

Exercice 3.7

(***) [Moindres carrés généralisés] On suppose le modèle linéaire général (3.5), mais seulement sous le postulat **P1** (erreurs centrées). Cependant, on suppose également connue la matrice de variance-covariance des erreurs (qui n'est donc pas forcément diagonale), qui sera notée Σ, que l'on suppose de rang n.

i. Déterminer alors l'espérance et la matrice de variance-covariance de $\widehat{\theta}$, estimateur des moindres carrés que l'on appeler ici moindres carrés ordinaires.

ii. On considère maintenant, en lieu et place de la distance euclidienne classique dans $\mathbb{R}^n$ utilisée pour les moindres carrés ordinaires, la distance définie par la norme :
$$\|U - V\|_\Sigma = (U - V)' \cdot \Sigma^{-1} \cdot (U - V).$$
L'estimateur $\widehat{\theta}_G$ de θ par moindres carrés généralisés minimisera $\|Y - X \cdot \theta\|_\Sigma$. Montrer que $\widehat{\theta}_G = (X' \cdot \Sigma^{-1} \cdot X)^{-1} X' \cdot \Sigma^{-1} \cdot Y$, et en déduire son espérance et sa matrice de variance-covariance.

iii. En déduire que l'estimateur $\widehat{\theta}_G$ a une matrice de variance-covariance "plus petite" que $\widehat{\theta}$ (c'est-à-dire que la différence entre leur matrice de variance-covariance est une matrice semi-définie positive). Ce résultat est une variante du Théorème de Gauss-Markov vu à l'exercice 6 précédent

iv. En supposant de plus que les observations sont conjointement gaussiennes, montrer que $\widehat{\theta}_G$ est l'estimateur du maximum de vraisemblance.

Exercice 3.8

(***) [Statistique des diffusions en modèles de la finance] On considère un actif financier dont la valeur au temps t est notée X_t. Un modèle classique en mathématiques financières consiste à supposer que le logarithme de cette valeur, $Z_t := log(X_t)$, suit un modèle de diffusion. Dans sa version la plus simple (drift constant et variance infinitésimale constante), ce modèle s'écrit :

$$dZ_t = \theta \cdot dt + \sigma \cdot dW_t, \tag{3.7}$$

où θ et $\sigma > 0$ sont des paramètres réels inconnus et W_t est un mouvement brownien, c'est-à-dire un processus continu gaussien centré à accroissement stationnaires et indépendants, ce que l'on peut encore quantifier par le fait que $\forall (s,t) \in \mathbb{R}_+^2$, $\mathbb{E}(W_t) =$

0 et $\mathbb{E}(W_s.W_t) = \min(t,s)$). Dans notre cas particulier, l'équation différentielle stochastique (3.7) a une solution très simple :

$$Z_t - Z_0 = \theta \cdot t + \sigma \cdot W_t. \tag{3.8}$$

Des propriétés du mouvement brownien, nous n'utiliserons que les deux suivantes :

- $W_0 = 0$;
- les accroissements de $(W_t)_t$ vérifient la propriété suivante : pour tout $n \in \mathbb{N}$ et pour tout $t_0 = 0 < t_1 < \cdots < t_n \in \mathbb{R}_+^n$, les variables $W_{t_1}, \big(W_{t_2} - W_{t_1}\big), \cdots, \big(W_{t_n} - W_{t_{n-1}}\big)$ sont des variables aléatoires indépendantes de lois respectives $\mathcal{N}(0,t_1), \mathcal{N}(0,t_2-t_1), \cdots, \mathcal{N}(0,t_n-t_{n-1})$.

Comme X_t (ou Z_t) ne peut être observé en temps continu, on suppose que l'observation est constituée de n points $(X_{t_i})_{1\le i\le n}$ uniformément répartis en temps, soit $t_i = i\Delta$, $i = 1, \cdots, n$ pour un certain pas de discrétisation $\Delta > 0$. On posera de plus par convention $t_0 = 0$.

i. En posant
$$Y_i := Z_{t_i} - Z_{t_{i-1}}, \quad i = 1, \cdots, n,$$
montrer que le vecteur Y suit un modèle linéaire et que les estimateurs optimaux de θ et σ sont
$$\begin{aligned} \widehat{\theta} &= \frac{Z_{t_n} - Z_{t_0}}{n\Delta} \\ \widehat{\sigma^2} &= \frac{1}{(n-1)\Delta} \sum_{i=1,n} \left(\big(Z_{t_i} - Z_{t_{i-1}}\big) - \frac{Z_{t_n} - Z_{t_0}}{n} \right)^2 \end{aligned}$$

ii. Montrer que ces estimateurs convergent quand n tend vers l'infini, Δ restant fixé.

Chapitre 4

Problèmes spécifiques à la régression

Dans ce chapitre, après quelques propos généraux sur la régression linéaire et non linéaire, nous allons essayer principalement de répondre à deux questions :

- *quel genre de phénomène peut être modélisé par un modèle de régression ?*
- *comment modifier un modèle de régression pour qu'il s'adapte mieux aux données ?*

1 Modèles linéaires et non linéaires

Nous avons vu que la régression multiple est un modèle linéaire pouvant s'écrire comme (3.5). Parmi les variables explicatives (on dit aussi régresseurs dans ce cas), certaines peuvent être intimement liées : la régression polynômiale décrite ci-dessous en est un tel exemple :

Soit Y la variable à expliquer et Z une variable explicative pertinente mais dont l'effet n'est pas linéaire. On propose le modèle polynomial suivant,

$$Y_i = \mu + \beta_1 Z_i + \beta_2 Z_i^2 + \beta_3 Z_i^3 + \beta_4 Z_i^4 + \varepsilon_i \quad \text{pour} \quad i = 1, \cdots, n.$$

C'est un modèle de régression multiple avec 4 régresseurs qui sont les puissances successives de la variable explicative. Il s'écrit encore sous la forme d'un modèle linéaire

général (3.5) avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & Z_1 & Z_1^2 & Z_1^3 & Z_1^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Z_n & Z_n^2 & Z_n^3 & Z_n^4 \end{pmatrix} \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \ . \end{pmatrix} = X \cdot \theta + \varepsilon$$

Ce modèle est polynomial en Z_i, mais linéaire en les paramètres inconnus μ, β_1, β_2, β_3, β_4.

De la même façon, le modèle suivant, appelé modèle de *régression périodique*, est un modèle linéaire statistique :

$$Y_i = \mu + \beta_1 Z_i + \beta_2 Z_i^2 + \gamma_1 \sin Z_i + \gamma_2 \sin 2Z_i + \gamma_3 \cos Z_i + \gamma_4 \cos 2Z_i + \varepsilon_i \ .$$

Il existe aussi des *modèles non-linéaires* (voir en particulier Tomassone *et al.* 1983, chap 5), par exemple :

- les *modèles exponentiels*, issus des *modèles à compartiments*, soit par exemple,

  $$Y_i = \beta_1 \exp(-\alpha_1 t_i) + \beta_2 \exp(-\alpha_2 t_i) + \varepsilon_i \quad \text{pour} \quad i = 1, \cdots, n.$$

  Les paramètres inconnus sont $\alpha_1, \alpha_2, \beta_1, \beta_2$ et la dépendance en α_1, α_2 est non linéaire à cause de l'exponentielle ;

- les *modèles logistiques*, soit par exemple,

  $$Y_i = \frac{\beta_1 + \beta_2 \exp(\beta_3 x_i)}{1 + \beta_4 \exp(\beta_3 x_i)} + \varepsilon_i \quad \text{pour} \quad i = 1, \cdots, n.$$

  Les paramètres inconnus sont $\beta_1, \beta_2, \beta_3, \beta_4$ et la dépendance est encore non linéaire.

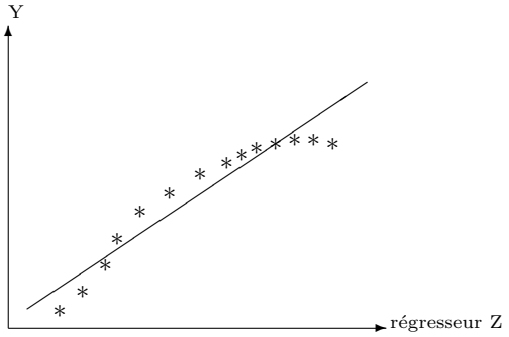
Le traitement numérique et statistique de tels modèles non-linéaires est considérablement plus délicat que celui des modèles linéaires. C'est pour cela que **l'on commencera en général dans toute approche de modélisation par le modèle linéaire dont on vérifiera la légitimité avant de passer à tout autre modèle.**

2 Contrôle graphique a posteriori

Lorsque l'on a posé un modèle de régression il est indispensable de commencer par s'entourer de "protections" graphiques pour vérifier empiriquement les 4 postulats de

base (ou tout au moins les postulats **P1-3**, puisque le postulat **P4** n'est pas vraiment important dès que l'on dispose de suffisamment de données).

• En régression linéaire simple, la confrontation graphique entre le nuage de points (z_i, y_i) et la droite de régression de Y par Z par moindres carrés ordinaires donne une information quasi exhaustive. En voici un exemple :



Sur ce graphique, on voit une courbure de la "vraie" courbe de régression de Y et on peut penser que le modèle est inadéquat et que le premier postulat **P1** n'est pas vérifié.

• Dans le cas de la régression multiple, ce type de graphique n'est pas utilisable car il y a plusieurs régresseurs. Les différents postulats sont à vérifier sur les termes d'erreur ε_i qui sont malheureusement inobservables. On utilise leurs prédicteurs naturels, les résidus : $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$. Par exemple, pour le modèle général de régression,

$$\begin{aligned} Y_i &= \mu + \beta_1 Z_i^{(1)} + \cdots + \beta_p Z_i^{(p)} + \varepsilon_i, \quad \text{pour} \;\; i = 1, \cdots, n, \\ &\Longrightarrow \quad \widehat{\varepsilon}_i = Y_i - \widehat{\mu} - \widehat{\beta}_1 Z_i^{(1)} - \widehat{\beta}_2 Z_i^{(2)} - \cdots - \widehat{\beta}_p Z_i^{(p)} \quad \text{pour} \;\; i = 1, \cdots, n. \end{aligned}$$

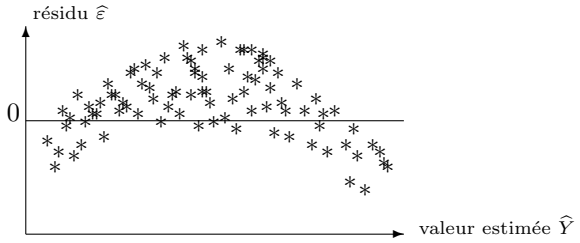
Voici maintenant plusieurs démarches permettant de s'assurer de la légitimité des conclusions, démarches à effectuer pour toute régression linéaire multiple :

1/ Pour vérifier les postulats P1 et P2 : adéquation et homoscédasticité

Le graphique le plus classique consiste à représenter les résidus $(\widehat{\varepsilon}_i)_i$ en fonction des valeurs prédites $(\widehat{Y}_i)_i$. Ce graphique doit être fait pratiquement systématiquement. Cela revient encore à tracer les coordonnées du vecteur $P_{[X]^\perp}.Y$ en fonction de celles de $P_{[X]}.Y$. L'intérêt d'un tel graphe réside dans le fait que si les quatre postulats **P1-4** sont bien respectés, il y a indépendance entre ces deux vecteurs qui sont centrés et

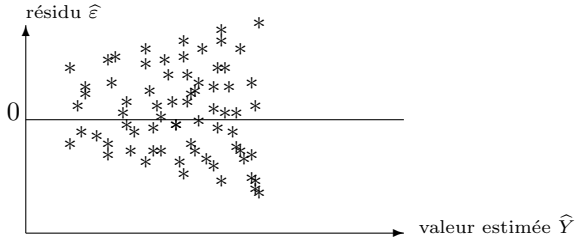
gaussiens (d'après le Théorème de Cochran). Cependant, à partir de ce graphe, on ne pourra s'apercevoir que de la possible déficience des postulats **P1** et **P2**, les deux autres postulats pouvant être "contrôlés" par d'autres représentations graphiques (voir plus loin). Concrètement, si on ne voit rien de notable sur le graphique (c'est-à-dire que l'on observe un nuage de points centré et aligné quelconque), c'est très bon signe : les résidus ne semblent alors n'avoir aucune propriété intéressante et c'est bien ce que l'on demande à l'erreur. Voyons justement maintenant deux types de graphes résidus/valeurs prédites "pathologiques" :

Type 1 "forme banane"



Dans ce cas on peut penser que le modèle n'est pas adapté aux données. En effet, il ne semble pas y avoir indépendance entre les $\widehat{\varepsilon}_i$ et les $\widehat{Y}_i$ (puisque, par exemple, les $\widehat{\varepsilon}_i$ ont tendance à croître lorsque les $\widehat{Y}_i$ sont dans un certain intervalle et croissent). Il faut donc améliorer l'analyse du problème pour proposer d'autres régresseurs pertinents, ou transformer les régresseurs $Z^{(i)}$ par une fonction de type $(\log, \sin)$, ce que l'on peut faire sans précautions particulières.

Type 2 "forme trompette"



Dans ce cas la variance des résidus semble inhomogène, puisque les $\widehat{\varepsilon}_i$ ont une dispersion de plus en plus importante au fur et à mesure que les $\widehat{Y}_i$ croissent. Un changement de variable pour Y pourrait être une solution envisageable pour "rendre" constante la variance du bruit (voir un peu plus bas)..

Remarque : certaines options sophistiquées utilisent plutôt des résidus réduits (Studentised residuals) qui sont ces mêmes résidus divisés par un estimateur de leur écart-type (généralement l'écart-type empirique) : cela donne une information supplémentaire sur la distribution des résidus qui doit suivre alors (toujours sous les postulats **P1-4**) une loi de Student. Cependant, on perd en capacité d'interprétation car le résidu est "adimensionnel", il n'est plus exprimé dans les unités de départ. Supposons par exemple que l'on veuille modéliser la taille (stature) d'adultes mesurée en mm. Un résidu de 5 correspond a une erreur de $5mm$ ce qui est tout-à-fait négligeable en pratique. Un résidu réduit est le plus souvent entre -2 et 2 (domaine de variation de la loi normale) sa valeur n'est pas directement interprétable.

Modifications possibles à apporter au modèle :

• On peut librement transformer les régresseurs $Z^{(1)}, \cdots, Z^{(p)}$ par toutes les transformations algébriques ou analytiques connues (fonctions puissances, exponentielles, logarithmiques,...), pourvu que le nouveau modèle reste interprétable. Cela peut permettre d'améliorer l'adéquation du modèle ou de diminuer son nombre de termes si on utilise ensuite une procédure de choix de modèles comme décrit au chapitre 9.

• En revanche, on ne peut envisager de transformer Y, que si les graphiques font suspecter une hétéroscédasticité. Dans ce cas, cette transformation doit obéir à des règles précises basées sur la relation suspectée entre l'écart-type résiduel σ et la réponse Y : c'est ce que précise le Tableau 2. Souvent ces situations correspondent à des modèles précis. Par exemple, la cinquième transformation correspond le plus souvent à des données de comptage. Dans le cas où les effectifs observés sont faibles (de l'ordre de la dizaine), on aura plutôt intérêt à utiliser un modèle plus précis basé sur des lois binomiales. Il s'agit alors d'un modèle linéaire généralisé. D'ailleurs toutes les situations issues d'une des transformations ci-dessus peuvent être traitées par modèle linéaire généralisé. Il n'entre pas dans le champ de ce cours de préciser ces modèles (on pourra consulter par exemple le livre de McCullagh et Nelder [44]).

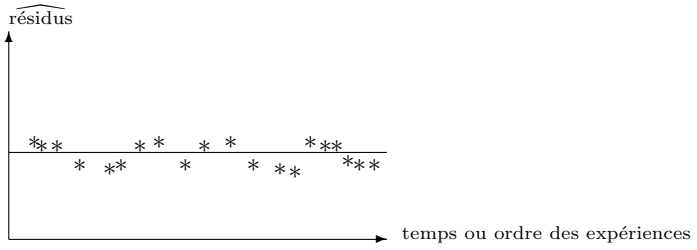
Notons cependant que pour des grands échantillons la transformation de Y peut suffire à transformer le modèle en un modèle linéaire classique et est beaucoup plus simple à mettre en œuvre. Par exemple, dans une étude bactériologique sur des désinfectants dentaires, on a mesuré le degré d'infection d'une racine dentaire en comptant les germes au microscope électronique. Sur les dents infectées, le nombre de germes est élevé et variable. L'écart-type est proportionnel à la racine carrée de la réponse. Une loi ayant cette propriété est la loi de Poisson, qui donne alors lieu à un modèle linéaire généralisé. Toutefois, si les décomptes sont en nombre important, travailler directement avec pour donnée la racine carrée du nombre de germe peut répondre tout aussi bien à la question.

Nature de la relation	Domaine pour Y	Transformation
$\sigma = (\text{cte})Y^k$, $k \neq 1$	$\mathbb{R}_+^*$	$Y \mapsto Y^{1-k}$
$\sigma = (\text{cte})\sqrt{Y}$	$\mathbb{R}_+^*$	$Y \mapsto \sqrt{Y}$
$\sigma = (\text{cte})Y$	$\mathbb{R}_+^*$	$Y \mapsto \log Y$
$\sigma = (\text{cte})Y^2$	$\mathbb{R}_+^*$	$Y \mapsto Y^{-1}$
$\sigma = (\text{cte})\sqrt{Y(1-Y)}$	$[0,1]$	$Y \mapsto \arcsin\left(\sqrt{Y}\right)$
$\sigma = (\text{cte})\sqrt{1-Y}.Y^{-1}$	$[0,1]$	$Y \mapsto (1-Y)^{1/2} - 1/3(1-Y)^{3/2}$
$\sigma = (\text{cte})(1-Y^2)^{-2}$	$[-1,1]$	$Y \mapsto \log(1+Y) - \log(1-Y)$

TABLE 4.1 – Table des changements de variable pour la variable à expliquer

2/ Pour vérifier le postulat P3 : indépendance

Un graphe pertinent pour s'assurer de l'indépendance des résidus entre eux et celui des résidus estimés $\widehat{\varepsilon}_i$ en fonction de l'ordre des données (lorsque celui-ci a un sens, en particulier s'il représente le temps). Par exemple, on peut obtenir le graphe suivant :



Un graphique comme celui ci-dessus est potentiellement suspect car les résidus ont tendance à rester par paquets lorsqu'ils se trouvent d'un côté ou de l'autre de 0. On pourra confirmer ces doutes en faisant un test de runs. Ce test est basé sur le nombre de runs, c'est-à-dire sur le nombre de paquets de résidus consécutifs de même signe. Sur le graphique ci-dessus, il y a 8 runs. On trouve les références de ce test dans tout ouvrage de tests non-paramétriques ou dans un livre comme celui de Draper et Smith [24] (p. 157). Voir également l'exercice 6.

Par ailleurs, si les erreurs sont corrélées suivant certaines conditions (par exemple si ce sont des processus ARMA), il est tout d'abord possible d'obtenir encore des résultats quand à l'estimation des paramètres, mais il existe également des méthodes de correction (on peut penser par exemple à des estimations par moindres carrés généralisés ou pseudo-généralisés ; voir par exemple Amemiya [3], Green [30], Guyon

[31] ou Jobson [35]).

3/ Pour vérifier le postulat P4 : gaussianité

Nous l'avons déjà évoqué et nous le redirons et le montrerons : le postulat **P4** de gaussianité des données n'est important que si l'on dispose de très peu de données (c'est-à-dire grossièrement, car tout dépend du modèle et du nombre de variables, moins de quelques dizaines). Dans ce cas, notamment pour que les tests de Fisher et de Student aient un sens, il peut être intéressant de vérifier si ce postulat peut être acceptable. Pour cela, **nous déconseillons fortement les tests d'adéquation classiques de Kolmogorov-Smirnov, Cramer-Von Mises,**..., du fait qu'on les appliquera sur les résidus $\widehat{\varepsilon}_i$, qui ne sont (quasiment) jamais indépendants. On préférera **se "contenter" d'une vérification graphique à partir du tracé d'une droite de Henri, (dite encore graphique QQ-plot)**. Celle-ci relie les points de $\mathbb{R}^2$ formés par les quantiles empiriques des résidus studentisés (c'est-à-dire les $\widehat{\varepsilon}_i$ divisés par leur écart-type empirique) en fonction des quantiles théoriques (pour les probabilités $k/(n+1)$ où $k = 1, \cdots, n$, le nombre de données étant n) d'une loi normale centrée réduite. La loi de Student "ressemblant" fortement à une loi gaussienne dès que le paramètre dépasse la dizaine, si les erreurs (ε_i) sont gaussiennes, c'est-à-dire si le postulat **P4** est vrai, alors la droite de Henri est une bissectrice du plan. Ce type de tracé permet surtout de voir si une loi "à queue de distribution lourde" ne pourrait être plus adéquate (dans ce cas, les points s'éloignent de la droite de Henri en ses extrémités).

3 Trouver la bonne régression

3.1 Erreur sur les régresseurs

La théorie du modèle de régression telle que nous l'avons vue dans le chapitre précédent, suppose que Y est aléatoire et que les régresseurs $Z^{(1)}, Z^{(2)}, \cdots, Z^{(p)}$ sont connus et non aléatoires. Les résultats peuvent s'étendre par conditionnement au cas où les régresseurs sont aléatoires, à condition qu'ils soient indépendants des erreurs ε. Il y a alors deux situations à considérer. Pour en donner une illustration, considérons l'exemple d'un phénomène chimique dépendant de la température T d'un bain suivant une fonction connue. Deux cas sont alors possibles :

- **Cas 1 :** On ne sait pas bien fixer la température du bain mais on sait très bien la mesurer. On avait prévu une expérience à 10^0C, 20^0C et 30^0C mais, en fait, on a effectivement des températures de 11^0C, 20^0C et 28^0C. Il est clair que la vraie variable pertinente est la température réelle et non pas celle planifiée. Si on introduit cette dernière dans la régression, on obtiendra un modèle parfaite-

ment correct, bien que la température réelle ait un certain caractère aléatoire.

– **Cas 2 :** On ne sait ni fixer ni mesurer précisément la température. Dans ce cas, on ne possède pas un régresseur vraiment pertinent (la vraie température) et en toute rigueur, on n'a pas le droit de poser un modèle de régression. Il faudrait plutôt utiliser des méthodes multivariées du type analyse en composantes principales , voir [52]. Cependant, en pratique, on acceptera de faire quand même une régression si l'erreur de mesure est plus faible que l'erreur d'ajustement du modèle.

3.2 Un cas particulier de régression : l'étalonnage ou calibration

Soit le problème suivant :

On considère la variable Z représentant l'âge d'un fœtus et la variable T, la longueur du fémur mesurée par échographie. On veut prédire Z à partir de T. Pour cela, on réalise une première expérience dite d'étalonnage : on constitue un échantillon de femmes enceintes à cycle menstruel régulier, ce qui permet de connaître précisément la date de conception et donc l'âge du fœtus. On mesure alors à l'aide d'une échographie la longueur des fémurs de leur fœtus. Dans la deuxième expérience dite de prédiction, on utilise l'expérience précédente pour prédire l'âge d'autres fœtus à partir de la mesure de la longueur fémorale.

Quel modèle utiliser ? Une première réponse, trop naïve, est la suivante :

La variable à expliquer est l'âge Z et la variable explicative est la longueur fémorale T.

Bien que l'on puisse adopter ce point de vue sous certaines conditions, il faut bien être conscient que ce n'est pas la démarche normale. En effet dans l'expérience d'étalonnage l'âge est parfaitement connu. Par contre pour un âge donné, les longueurs fémorales se répartissent autour d'une valeur moyenne, à cause de la variabilité génétique de la population humaine considérée. Pour respecter les postulats **P1-4** du modèle linéaire il vaut mieux écrire

$$\text{longueur du fémur} = T = \alpha + \beta \cdot Z + \varepsilon.$$

Dans la deuxième phase, l'équation de régression sera inversée. En effet

$$T = \alpha + \beta \cdot Z \Longleftrightarrow Z = \frac{T - \alpha}{\beta}$$

(lorsque la première phase d'e'talonnage est significative, c'est-à-dire $\beta \neq 0$), on peut dans une seconde phase estimer l'âge Z à partir de la longueur T du fémur. L'intérêt

du cas particulier de l'étalonnage est de bien illustrer ce que veut dire un modèle de régression.

3.3 Choisir parmi les régresseurs

Dans de nombreux problèmes concrets, on ne désire pas conserver tous les régresseurs, mais plutôt éliminer tous ceux qui n'apportent pas d'explication supplémentaire pour la variable à expliquer Y.

La régression avec l'ensemble des régresseurs donne un test de Student portant sur la nullité des coefficients de chaque variable. Ce test correspond à l'hypothèse d'enlever une variable en conservant toutes les autres. On se heurte alors au problème suivant : dans le cas général, dès que l'on enlève une variable, tous les tests de Student des autres variables sont modifiés.

La difficulté pour éliminer certaines variables explicatives (ou régresseurs) vient le plus souvent de leur possible colinéarité. Imaginons que deux variables $Z^{(1)}$ et $Z^{(2)}$ soient "presque" colinéaires et très liées avec la variable à expliquer Y.

- dans le modèle avec $Z^{(1)}$ seul, $Z^{(1)}$ est significatif ;
- dans le modèle avec $Z^{(2)}$ seul, $Z^{(2)}$ est significatif ;
- dans le modèle avec les deux variables, aucune des variables explicatives n'est significative.

Dans une telle situation, les deux modèles (avec $Z^{(1)}$ seul ou avec $Z^{(2)}$ seul) conviennent et il sera très difficile de choisir entre la première et la seconde possibilité.

Trouver le meilleur ensemble de régresseurs sera difficile pour deux raisons. Tout d'abord, si l'on considère un total de k régresseurs, il y 2^k sous-modèles possibles. Il est courant de manipuler des modèles avec une vingtaine de régresseurs, ce qui correspond à environ un millon de sous-modèles possibles. Cela représente une importante complexité numérique. Ensuite, s'il y a une colinéarité entre différents régresseurs, plusieurs sous-modèles peuvent donner des résultats identiques. Aussi, va-t-on commencer par mesurer la colinéarité entre les régresseurs, ce qui peut être fait à partir des deux étapes suivantes :

i. en premier lieu, qualitativement, on peut observer si les tests de Student ou de Fisher de nullité des coefficients des différentes variables sont modifiés suivant que l'on change de modèle ;

ii. ensuite, et si l'étape précédente conduit au soupçon de colinéarité, on mesure quantitativement la "dépendance linéaire" d'un régresseur par rapport aux autres par le VIF (Variance Inflation Factor, facteur d'augmentation de la variance). Soit $Z^{(i)}$ le i-ème régresseur parmi k autres régresseurs. On peut alors effectuer une régression de $Z^{(i)}$ sur les autres régresseurs. On calcule alors le coefficient de détermination R_i^2 (pourcentage de variance expliquée) de $Z^{(i)}$ régressé sur les autres régresseurs. Soit alors $\widehat{Z}^{(i)}$ l'estimation de $Z^{(i)}$ par une telle régression. On a donc :

$$R_i^2 = \frac{\|\widehat{Z}^{(i)} - \overline{Z}^{(i)}\|^2}{\|Z^{(i)} - \overline{Z}^{(i)}\|^2} \text{ où } \overline{Z}^{(i)} \text{ est obtenu par régression de } Z^{(i)} \text{ par les } Z^{(j)},\ j \neq i.$$

Ce coefficient est aussi le rapport de la variance résiduelle par la variance totale. On définit ainsi le VIF associé à $Z^{(i)}$ par l'expression :

$$VIF_i = \frac{1}{1 - R_i^2}.$$

Ce coefficient VIF est une quantité toujours supérieure à 1. Elle vaut 1 quand le régresseur n'est pas du tout colinéaire aux autres régresseurs (il est orthogonal, au sens de l'algèbre linéaire) car alors $R_i^2 = 0$. Une valeur supérieure à 10 est considérée comme un signe de colinéarité importante (ceci ne donne qu'une idée imprécise de la situation : cette valeur de 10 est à moduler en fonction du nombre de données et du niveau de confiance qu'un test sur le VIF demanderait).

Remarque : Certains cours sur le modèle linéaire définissent également le coefficient TOL (tolerance) qui se définit comme l'inverse du VIF : $TOL_i = \frac{1}{VIF_i} = 1 - R_i^2$. On conçoit bien la similitude des résultats obtenus avec l'un ou l'autre des indices.

Nous finirons ce chapitre par des remarques générales concernant l'heuristique d'une modélisation. Pour commencer, nous dirons qu'un modèle de régression peut être explicatif ou prédictif :

- un modèle est explicatif quand il y a une vraie liaison causale (par exemple issue d'une loi physique ou chimique) entre la variable à expliquer et les régresseurs.

- un modèle est seulement prédictif quand il n'a pas la propriété précédente, mais prédit bien la variable à expliquer. Un autre modèle pourrait, a priori, tout aussi bien convenir.

Toute personne ayant l'habitude de la statistique sait que l'on n'est jamais sûr d'obtenir un modèle explicatif. Par exemple on peut observer dans certains cas extrêmes une

liaison négative entre (a) le rendement d'une parcelle et (b) son taux de traitement par produits physio-sanitaires. L'explication de ce paradoxe vient de ce que l'on ne traite que quand la parcelle est infectée et donc quand le rendement est bas. La liaison n'est pas causale, car arrêter le traitement n'augmentera pas le rendement, bien au contraire.

Dans la recherche d'un modèle, on distingue deux solutions extrêmes entre lesquelles la réalité se trouvera le plus souvent :

- le bon cas : les VIF sont faibles et le test de Student d'une variable dans n'importe quel sous-modèle est qualitativement le même. Dans ce cas, on aura l'espoir de trouver un modèle explicatif ;

- le mauvais cas : certains VIF sont élevés et les tests de Student d'une variable changent suivant les variables associées. Dans ce cas, on ne pourra qu'escompter trouver un modèle prédictif.

4 Stratégies de sélection d'un modèle explicatif

Les études précédentes du modèle linéaire nous offre un premier choix de stratégies pour essayer de sélectionner un modèle explicatif par rapport à un autre, ou par rapport à tous les autres. En effet, nous avons vu des techniques pour tester la validité d'un sous-modèle par rapport à un modèle plus "grand". Les critères qui vont guider notre choix sont au nombre de trois :

– le pourcentage de variance expliquée par un sous-modèle, appelé encore le coefficient de détermination :

$$R^2 = \frac{\text{SC totale} - \text{SCR}}{\text{SC totale}} = \frac{\|\widehat{Y}^{(0)} - \overline{Y}\|^2}{\|Y - \overline{Y}\|^2},$$

où l'estimation $\widehat{Y}^{(0)}$ est obtenue pour le sous-modèle considéré ;

– le carré moyen résiduel : $\widehat{\sigma}_n = \dfrac{1}{n - k_0}\|Y - \widehat{Y}^{(0)}\|^2$, où l'estimation $\widehat{Y}^{(0)}$ est obtenue pour le sous-modèle, qui est de dimension k_0 ;

– la comparaison de deux modèles emboîtés par la statistique du test F.

Notons que **le premier critère, utilisant le coefficient de détermination, privilégie toujours les grands modèles, dans le sens où plus on ajoute de**

régresseurs, plus la somme des carrés résiduelle $\|Y - \widehat{Y}^{(0)}\|^2$ diminue, donc plus le coefficient R^2 se rapproche de 1. Il ne permet de comparer que des modèles de même taille. Par exemple, si on cherche un modèle polynômial représentant des données, il est bien clair que plus on augmente le degré du polynôme, plus on améliore l'adéquation du modèle aux données. A la limite, un modèle polynômiale ayant pour degré le nombre de données (moins un) sera totalement en adéquation ; pourtant il ne sera certainement pas le plus intéressant...

On suppose donc maintenant que l'on dispose d'un certain nombre de variables explicatives et que seules certaines d'entre elles interviennent dans le vrai modèle. On suppose de plus que ces variables sont peu corrélées entre elles, ou plus précisément que les différents coefficients VIF sont proches de 1. Voyons alors les stratégies possibles pour tenter d'identifier le vrai modèle.

- **Stratégie 1** : *régression descendante* (conseillée au débutant). On pose le grand modèle (celui avec tous les régresseurs possibles) et ensuite, à chaque étape, on calcule la statistique F correspondant au retrait de chaque variable (nullité du coefficient associé). On enlève du modèle la variable associée au F de plus faible valeur. On arrête cette procédure quand la valeur du F de la variable enlevée dépasse un certain quantile lié au niveau du test fixé par l'utilisateur (par exemple la valeur critique au niveau 5% pour laquelle on aura une P-value inférieure à 0.05).

- **Stratégie 2** : *régression ascendante* (déconseillée). On part d'un petit modèle censé représenter relativement bien les données, et à chaque étape on rajoute parmi les régresseurs non utilisés celui qui a le F d'introduction le plus élevé. En d'autres termes, on ajoute le régresseur le plus pertinent, celui qui améliore par exemple le coefficient R^2. On s'arrête quand le F de la variable introduite est inférieur à une valeur donnée par l'utilisateur (qui peut être la valeur critique à 5%). L'inconvénient de cette stratégie est d'être mal fondée théoriquement : en effet, si on rajoute des régresseurs aux différents modèles, cela implique que tous les modèles sauf le dernier sont faux et donc que tous les calculs faits dans la stratégie ascendante ne sont pas strictement corrects.

- **Stratégie 3** : il existe une stratégie dite régression hiérarchique qui mélange les deux ci-dessus : à chaque étape, on peut ajouter ou enlever un régresseur. Nous ne détaillerons pas.

- **Stratégie 4** : *Algorithme de Furnival et Wilson* [29] . Cet algorithme très puissant permet de trouver "la meilleure régression" à p régresseurs parmi les k régresseurs possibles. Précisons que ce résultat n'est absolument pas garanti par les autres techniques, qui ne travaillent que sur certains sous-modèles. L'algo-

rithme donne la meilleure régression à 1 régresseur, puis à 2 régresseurs, puis à 3 régresseurs etc ... Il reste à fixer p et cela peut être obtenu par l'utilisation de critères à minimiser (voir chapitre 9). Cet algorithme est mis en œuvre dans Splus et SAS.

Ceci concerne donc la recherche d'un modèle explicatif. Cependant, en pratique, ce sera surtout en vue de nouvelles prédictions que l'on cherchera à établir un modèle. Dans ce cadre, on peut se contenter de sélectionner un "bon" modèle prédictif : le chapitre 9 est ainsi consacré à définir des critères de sélection de modèles prédictifs.

5 Exemple traité par logiciels informatiques

Nous allons utiliser dans ce chapitre (et d'autres qui suivent) un jeu de donnée issu de Tomassone *et al* [58]. Il s'agit d'une étude réalisée en 1973 sur un parasite du pin bien connu : la chenille processionnaire . On désire connaître l'influence de certaines variables sur la densité de peuplement du parasite. La variable à expliquer est notée $X11$: c'est le nombre moyen de nids par arbre sur la parcelle considérée de 10 hectares. On dispose ainsi au total des résultats concernant 32 parcelles distinctes. Les différentes variables susceptibles d'avoir une influence sur $X11$ (ou encore sur $X12 = \log(X11)$, qui s'avérera plus pertinente), sont relatives aux différentes caractéristiques de la placette (subdivision de chaque parcelle) et sont :

- l'altitude en mètres : $X1$;
- la pente en degrés : $X2$;
- le nombre de pins dans la placette : $X3$;
- la hauteur (en m) de l'arbre échantillonné au centre de la placette : $X4$;
- le diamètre de cet arbre : $X5$;
- la note de densité de peuplement : $X6$;
- l'orientation de la placette (de 1=sud, à 2=autre) : $X7$;
- la hauteur en mètres des arbres dominants : $X8$;
- le nombre de strates de végétation : $X9$;
- le mélange du peuplement (de 1=mélangé, à 2=non mélangé) : $X10$.

Il s'agit de données quantitatives même pour $X7$ ou $X9$, la valeur dans le fichier étant en fait une moyenne sur un certain nombre de placettes échantillonnées dans chaque parcelle. Dans cet exemple, on va étudier la régression de la variable $X11$ par les variables $X1$, $X2$, $X4$ et $X5$ dans le cadre d'un modèle linéaire (pour ne pas surcharger les résultats, nous avons donc délibérément choisi parmi les régresseurs, choix qui sera expliqué au chapitre 9). Les conclusions tirées des sorties numériques et graphiques seront présentées en fin de section.

Logiciel Splus :

Commençons par le logiciel Splus. La régression linéaire multiple de $X11$ par les variables $X1$, $X2$, $X4$ et $X5$, s'écrit de la manière suivante (on suppose ici que l'on a déjà construit la table `data.frame proc` enregistrée dans le `.Data`, répertoire contenant, sous forme de fichiers, toutes les variables utilisées par le logiciel) :

```
menuLm(X11~X1+X2+X4+X5,proc,plotResidVsFit.p=T,plotQQ.p=T,
                                    predict.p=T,ci.p=T,se.p=T)
```

Commentaires :

- en plus des options par défaut précisant le modèle linéaire choisi, on a également demandé d'afficher les valeurs prédites $\widehat{X}11$ par l'option `predict.p=T`, ainsi que les écart-types empiriques `se.p=T` associés à chaque prédiction, et les intervalles de confiance à 95% sur ces prédictions. D'autres valeurs numériques auraient pu être affichées (les résidus, les corrélations,...).

- deux graphes, les résidus en fonction des valeurs prédites et les QQ-plot, dorénavant classiques, ont été demandés. D'autres graphes auraient également pu être tracés (les résidus studentisés en fonction des valeurs prédites, les distances de Cook,...) en précisant `=T` ("true") dans la commande `menuLm` au lieu de `=F` ("false") qui est considérée par défaut.

On obtient ainsi les résultats suivants (on n'a conservé ici qu'un extrait des sorties numériques) :

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  6.6031  1.0242     6.4469  0.0000
         X1 -0.0028  0.0008    -3.5963  0.0013
         X2 -0.0456  0.0135    -3.3919  0.0022
         X4 -0.7551  0.2159    -3.4973  0.0016
         X5  0.1685  0.0515     3.2689  0.0029

Residual standard error: 0.5336 on 27 degrees of freedom
Multiple R-Squared: 0.6303
F-statistic: 11.51 on 4 and 27 degrees of freedom, the p-value is 0.00001394
```

```
          fit    se.fit        LCL95      UCL95
1  1.69644928 0.1624598  1.363109280  2.0297893
2  1.26021928 0.1798720  0.891152472  1.6292861
3  1.36419786 0.2090220  0.935320207  1.7930755
4  1.08230917 0.1650649  0.743623996  1.4209943
5  0.33414024 0.1657843 -0.006021016  0.6743015
6  1.02553202 0.1083952  0.803123407  1.2479406
:      :           :           :           :
```

Logiciel R :

Voyons maintenant les commandes nécessaires en R pour obtenir un traitement équivalent (on se référera aux extraits de résultats précédents), avec également le tracé (voir Figure 5) de plusieurs graphes (ne pas s'intéresser à celui concernant les distances de Cook, que nous ne considérerons pas dans cet ouvrage) permettant de "diagnostiquer" les possibles problèmes (au sens d'une remise en question des postulats initiaux) d'une telle régression :

```
library(car)
lm.proc=lm(X11~X1+X2+X4+X5,proc)
Anova(lm.proc,type="III")
par(mfrow=c(2,2))
plot(lm.proc,las=1)
```

Logiciel SAS :

Terminons cette fois-ci avec le logiciel SAS. On suppose les données rangées dans le data `sasuser.proc` et on soumet le programme suivant :

```
proc reg data=sasuser.proc all;
model X11=X1 X2 X4 X5;
plot r.*p.;run;quit;
```

Voici un extrait de la sortie (qui est de taille assez conséquente car la commande `all` précisée dans la procédure `reg` fournit toutes les sorties numériques possibles de cette procédure) :

```
            Correlation
```

Variable	X1	X2	X4	X5	X11
X1	1.0000	0.0861	0.3211	0.2876	-0.5337
X2	0.0861	1.0000	0.1346	0.1175	-0.4647
X4	0.3211	0.1346	1.0000	0.9050	-0.3576
X5	0.2876	0.1175	0.9050	1.0000	-0.1578
X11	-0.5337	-0.4647	-0.3576	-0.1578	1.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	13.10487	3.27622	11.51	<.0001
Error	27	7.68670	0.28469		
Corrected Total	31	20.79157			

Root MSE	0.53357	R-Square	0.6303
Dependent Mean	0.81406	Adj R-Sq	0.5755
Coeff Var	65.54360		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	6.60309	1.02423	6.45	<.0001
X1	X1	1	-0.00281	0.00078216	-3.60	0.0013
X2	X2	1	-0.04565	0.01346	-3.39	0.0022
X4	X4	1	-0.75510	0.21591	-3.50	0.0016
X5	X5	1	0.16847	0.05154	3.27	0.0029

Parameter Estimates

Variable	Label	DF	Tolerance	Inflation	Variance 95% Confidence Limits	
Intercept	Intercept	1	.	0	4.50153	8.70464
X1	X1	1	0.89499	1.11733	-0.00442	-0.00121
X2	X2	1	0.97975	1.02067	-0.07326	-0.01803
X4	X4	1	0.17631	5.67193	-1.19810	-0.31209
X5	X5	1	0.18093	5.52696	0.06273	0.27422

Output Statistics

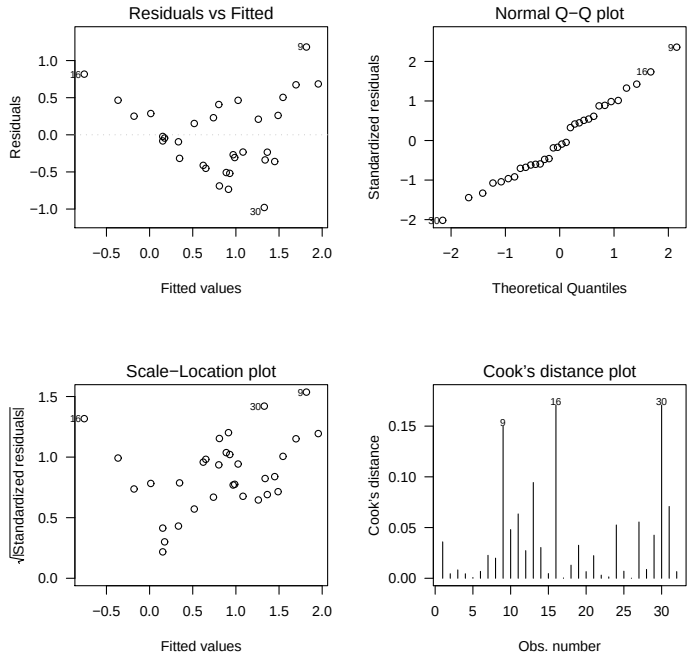


FIGURE 4.1 – Graphiques issus de la régression (R)

Dep Obs	Var X11	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		-2-1 0 1 2
1	2.3700	1.6964	0.1625	1.3631	2.0298	0.5520	2.8409	|**|
2	1.4700	1.2602	0.1799	0.8912	1.6293	0.1049	2.4155	| |
3	1.1300	1.3642	0.2090	0.9353	1.7931	0.1884	2.5400	| |
4	0.8500	1.0823	0.1651	0.7436	1.4210	-0.0637	2.2283	| |
5	0.2400	0.3341	0.1658	-0.0060	0.6743	-0.8123	1.4806	| |
6	1.4900	1.0255	0.1084	0.8031	1.2479	-0.0916	2.1427	|*|
7	0.3000	0.0136	0.2569	-0.5135	0.5407	-1.2015	1.2287	|*|
8	0.0700	-0.1807	0.2677	-0.7299	0.3686	-1.4055	1.0442	|*|
9	3.0000	1.8174	0.1836	1.4406	2.1942	0.6596	2.9752	|****|
10	1.2100	0.8020	0.2604	0.2677	1.3364	-0.4162	2.0203	|*|
:	:	:	:	:	:	:	:	: :

Conclusions issues des relevés des 3 logiciels : Pour l'explication de du nombre de nids $X11$, le test de Fisher global et les différents tests de Student sont significatifs pour les quatres variables considérées : $X1$, $X2$, $X4$ et $X5$. Le graphique des QQ-plot suggère que l'hypothèse de normalité a une certaine légitimité. En revanche, les deux graphiques de résidus (simples et studentisés) nous montrent plutôt des résidus dépendant des valeurs prédites. On remarque en particulier un triangle en bas à gauche où il n'y a pas de données. Cela est dû aux contraintes de positivité de l'observation qui sont mal gérées par notre modèle. Travailler avec la variable $X12 = \log(X11)$, permettra d'éviter ce genre de problème. Le résultat avec les signes des coefficients des différentes variables peut s'interpréter de la façon suivante : la chenille processionnaire semble être moins présente sur les parcelles difficiles d'accès, c'est-à-dire celles situées en altitude ($X1$ élevé) avec une pente forte ($X2$ élevé) et dont les arbres sont hauts ($X4$ élevé) et fins ($X5$).

6 Exercices

Exercice 4.1

(*) [Moindres carrés pondérés] On suppose que l'on ne veut pas accorder la même importance à toutes les observations dans un modèle de régression multiple (voir un exemple important de ceci dans l'exercice 6 sur le filtrage exponentiel). Pour simplifier, on se place tout de même sous les les postulats **P1-3** (erreurs centrées, homoscédasticité et indépendance des erreurs). Soit $(p_1, \ldots, p_n)$ les poids respectifs que l'on accorde aux différentes variables, avec $0 < p_i$ pour $i = 1, \ldots, n$ et $\sum p_i = n$. Soit Ω la matrice diagonale composée de $(p_1, \ldots, p_n)$ sur la diagonale. Tout comme pour les moindres carrés généralisés, on considérera la norme :

$$\|U - V\|_\Omega = (U - V)' \cdot \Omega^{-1} \cdot (U - V).$$

et $\widehat{\theta}_\Omega$, l'estimateur qui minimise $\|Y - X\theta\|_\Omega$ Déterminer l'expression de l'estimateur $\widehat{\theta}_\Omega$ de θ, son espérance et sa variance, ainsi que l'expression de l'estimateur $\widehat{\sigma}^2_\Omega$ de σ^2 la variance des erreurs.

Exercice 4.2

(*) Soit $Z_1, \ldots, Z_m$, m réels. On rappelle que le déterminant de Van-der-Monde, défini par :

$$\triangle = \begin{vmatrix} 1 & Z_1 & \ldots & Z_1^{m-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & Z_m & \ldots & Z_m^{m-1} \end{vmatrix} = \prod_{1 \leq i < j \leq m} (Z_i - Z_j)$$

ne s'annule que si deux des Z_i sont égaux. On considère le modèle de régression quadratique :

$$Y_i = \mu + \beta_1 \cdot Z_i + \beta_2 \cdot Z_i^2 + \varepsilon_i \quad \text{pour} \quad i = 1, \ldots, n > 2.$$

i. Montrer que si les Z_i ne prennent que deux valeurs distinctes, la matrice X du modèle n'est pas de plein rang.

ii. Montrer que si les Z_i ne prennent que trois valeurs distinctes au moins, la matrice X du modèle est de plein rang.

iii. Généraliser au cas de la régression polynômiale de degré quelconque.

Exercice 4.3

(**) [Prédiction et Filtrage exponentiel] On suppose un modèle de régression multiple sous les postulats **P1-3**, dont les différentes variables (explicatives et à expliquer) dépendent du temps (les individus sont par exemple des jours ou des années). On connaît donc $Y_1, \ldots, Y_n$ et on veut utiliser un modèle linéaire pour prédire Y_{n+1}, sachant que l'on connaît par un moyen ou un autre $Z_{n+1}^{(1)}, \ldots, Z_{n+1}^{(k)}$.

i. Par quoi peut-on naturellement approcher Y_{n+1} si l'on utilise les résultats des moindres carrés ? On appellera cette variable aléatoire prédiction. Quelle est la variance (théorique) de l'erreur commise ? Par quelle valeur approchera-t-on cette variance ?

ii. Le modèle des moindres carrés ordinaires accorde le même poids à tous les individus. Pour effectuer une prédiction à l'instant $n+1$, il est raisonnable de penser, surtout dans le cas fréquent où l'on n'est pas sûr du modèle choisi (voir à ce sujet le chapitre 9), que ce sont essentiellement les individus proches dans le temps de n qui comptent et que plus le passé est lointain moins les individus du passé importent. On peut ainsi accordé un poids qui décroît exponentiellement vite avec le passé (d'où le nom de filtrage exponentielle). En utilisant un paramètre $0 < \beta \leq 1$, on pourra utiliser un modèle de moindres carrés pondérés dont les poids pourront s'écrire $p_i = p \cdot \beta^{n-i}$, où $p > 0$ est un paramètre de normalisation. Déterminer p. Que se passe-t-il lorsque $\beta = 1$? Quelle est alors l'estimation de Y_{n+1} par filtrage exponentiel ? Quelle est sa variance et comment l'approcher ?

Remarque : le problème de cette méthode est le choix de la valeur de β ; en pratique on choisit souvent $\beta = 0.8$ ou 0.7, mais une procédure dite adaptative, c'est-à-dire estimant, à partir des données, une valeur optimale de β, pour un certain critère, peut également être mise en place.
Un tel filtre exponentiel peut aussi être utiliser pour ce que l'on appelle une régression locale dont le cadre est le suivant : on suppose que l'on dispose d'une série chronologique $(Y_i)_{i \in \mathbb{N}}$ possédant une tendance mais pas de composante saisonnière (voir exercice 6), et telle que $Y_1, \cdots, Y_n$ est connue. Pour estimer la tendance, en vue par exemple d'identifier le bruit de la série (ε_i) et pour s'adapter aux possibles variations très locales de cette tendance, on peut effectuer une régression locale qui consiste en

une suite de régressions linéaires simples (ou polynômiales de faible degré) effectuées à l'aide d'un filtre exponentiel dont le centre est chaque couple (i_0, Y_{i_0}) de l'ensemble des données ; cela revient à considérer les couples $(i_0, \widehat{Y}_{i_0})$ pour $i_0 = 1, \ldots, n$, obtenus par une régression par moindres carrés pondérés (de matrice diagonale de poids $\Omega_{i_0} = (p \cdot \beta^{-|i_0 - i|} \cdot \delta_{ij})_{ij}$). On appelle parfois une telle méthode régression dite LOESS. Rendu plus robuste par la prise en compte d'une matrice de poids adapté à chaque point, on obtient une régression dite LOWESS. Ces procédures, non-paramétriques, sont présentes dans les trois logiciels R, SAS et Splus, et le filtre exponentiel peut également être remplacé par d'autres fonctions régulières et décroissant suffisamment vite.

Exercice 4.4

(**) [Transformation de variables] Soit des observations Y_{ij} qui suivent le modèle suivant :

$$Y_{ij} = \mu_i + \varepsilon_{ij} \cdot \sqrt{\mu_i(1 - \mu_i)}, \quad \text{pour} \;\; i = 1, \cdots, I, \;\; j = 1, \cdots, J, \tag{4.1}$$

où les erreurs ε_{ij} vérifient les postulats habituels **P1-4**. Ce modèle correspond au 5ème cas du tableau 2. On pose

$$Z_{ij} = \arcsin(\sqrt{Y_{ij}}).$$

i. Écrire un développement limité à l'ordre 1 de la fonction $x \mapsto \arcsin(\sqrt{x})$ au point x_0.

ii. On admet que l'on peut négliger le reste : soit c'est une hypothèse que l'on assume, soit on suppose que σ^2 tend vers zéro. Dans ce dernier cas on utilise ce que l'on appelle la δ méthode ou Théorème de Slutsky (voir Van der Vaart [60] ou Dacunha-Castelle et Duflo [21] p. 91). Montrer que Z_{ij} suit un modèle d'analyse de la variance à un facteur.

iii. Montrer que si les Y_{ij} sont des données de comptage sur de grands effectifs, avec une probabilité de succès qui dépend de l'indice i on est approximativement dans la situation du modèle (4.1).

iv. Traiter de même tous les cas du tableau 2.

Exercice 4.5

(**) [Test de runs] Ce test est utilisé pour tester la présence ou non de corrélations dans les ε_i. On commence d'abord par le décrire dans le cas où l'on observe des variables aléatoires $Y_1, \ldots, Y_n$ dont on veut tester l'indépendance. On les suppose de

médiane zéro. On compte parmi $Y_1, \cdots, Y_n$ le nombre R de "paquets" (ou "runs") de (Y_i) consécutifs ayant le même signe.
Par exemple, si $Y_1, \ldots, Y_9 = (1.1, 1.3, -2, -1, 4.5, 1.6, -2.7, -1.3, 4)$, il y a 5 runs pour $n = 9$ données.

i. Montrer que si on suppose qu 'aucun des Y_i n'est nul, alors :

$$R = 1 + \sum_{i=1}^{n-1} \mathbb{1}_{Y_i Y_{i+1} < 0} := 1 + \sum_{i=1}^{n-1} Z_i$$

ii. On suppose que les Y_i sont indépendantes et de loi diffuse (c'est-à-dire absolument continue par rapport à la mesure de Lebesgue). Montrer que $\mathbb{E}(R) = \dfrac{n+1}{2}$,

iii. Montrer que si $|i-j| > 1$, Z_i et Z_j sont indépendantes. Montrer que Z_i et Z_{i+1} sont également indépendantes. En déduire $\mathrm{Var}(R)$.

iv. En utilisant le théorème de la limite centrale, construire pour des grands échantillons une statistique libre qui suit une loi normale centrée réduite sous l'hypothèse H_0 d'indépendance et qui tend vers $\pm\infty$ sous les alternatives H_1 d'intrication et de répulsion. Nous laissons au lecteur le soin de deviner le sens de ces deux derniers mots.

Remarque : Pour ce qui est du test de l'indépendance des erreurs dans un modèle linéaire, on appliquera le test de runs aux estimateurs $\widehat{\varepsilon_i}$ en négligeant leurs liaisons (toujours présentes, même sous l'hypothèse d'indépendance des ε_i) et en négligeant le fait que leur médiane n'est qu'approximativement nulle. Il existe d'autres versions de ce test sous des hypothèses d'échangeabilité (voir par exemple Lecoutre et Tassi, [39]).

Exercice 4.6

(**) [Tendance et composante saisonnière d'une série chronologique] On considère une suite de variables aléatoires $(Y_i)_{i\in\mathbb{N}}$ définies sur le même espace de probabilité et possédant un moment d'ordre 2 constant σ^2. L'indice i est relatif au temps et peut représenter des jours, des années,... Une telle suite (Y_i) est appelée série chronologique. On suppose également que pour tout $i \in \mathbb{N}$, on peut décomposer Y_i de façon unique sous la forme :

$$Y_i = t(i) + s(i) + \varepsilon_i$$

où s est une fonction de $\mathbb{R}$ dans $\mathbb{R}$, appelée composante saisonnière, périodique de période $T \in \mathbb{N}^*$ connue et telle que $\sum_{i=1}^{T} s(i) = 0$, t est une fonction de $\mathbb{R}$ dans $\mathbb{R}$ appelée tendance et $(\varepsilon_i)_{i\in\mathbb{N}}$ est une suite de variables centrées appelée bruit. On suppose que la tendance t s'écrit sous la forme $t(x) = a_1 \cdot g_1(x) + \ldots + a_k \cdot g_k(x)$ pour

$x \in \mathbb{R}$, où $k \in \mathbb{N}^*$ est connu, $g_1, \ldots, g_k$ sont des fonctions connues et $a_1, \ldots, a_k$ sont des réels inconnus. Enfin, on suppose que l'on connaît une trajectoire $(Y_1, \ldots, Y_{nT})$, où $n \in \mathbb{N}$.

i. On suppose que $s = 0$. En utilisant une régression linéaire par moindres carrés ordinaires, quels sont les estimateurs $\widehat{t}(x)$ pour $x \in \mathbb{R}$ et $\widehat{\sigma}_1^2$?

ii. On suppose maintenant que $t = 0$ et que pour $k \in \mathbb{N}$, si k est un multiple de $i \in \{1, \ldots, T\}$ alors $s(k) = s_i$ (et donc $\sum_{i=1}^{T} s_i = 0$). En utilisant une régression linéaire par moindres carrés ordinaires (attention, les T paramètres s_i sont liés linéairement...), donner l'expression de $\widehat{\sigma}_2^2$ et montrer que

$$\widehat{s}_i = \frac{1}{n} \sum_{k=1}^{n} Y_{i+(k-1)T} - \frac{1}{nT} \sum_{k=1}^{nT} Y_k \qquad \text{pour} \quad i = 1, \ldots, T.$$

iii. Dans le cas où la tendance et la composante saisonnière ne sont pas nulles, quels estimateurs proposeriez vous pour ces fonctions, ainsi que pour la variance ? (on suppose les mêmes écritures de t et s que précédemment). Ces estimateurs sont-ils les mêmes que ceux obtenus en estimant d'abord la tendance, et ensuite la composante saisonnière à partir de la série à laquelle on a retiré cette estimation de la tendance ?

Exercice 4.7

(***) [Interpolation et splines] On suppose que l'on dispose de n couples de réels (x_i, y_i), les x_i étant tous distincts les uns des autres, à partir desquels on désirerait avoir une idée d'une fonction $y_i = g(x_i)$.

i. Montrer qu'il existe un unique polynôme de degré $n-1$ passant par l'ensemble des points. On pourra utiliser l'exercice 6. Rappeler la formule de ce polynôme d'interpolation dit de Lagrange (indication : ce polynôme s'écrit comme une combinaison linéaire de polynômes de la forme $P_j(x) = \prod_{i \neq j} \frac{x - x_i}{x_j - x_i}$ pour $j = 1, \ldots, n$). Quelles sont les limitations d'une telle estimation de g ? Quelle régression linéaire suffirait-il de mettre en place pour obtenir le même polynôme ?

ii. Une autre méthode d'interpolation "déterministe" consiste en l'utilisation de splines cubiques, qui sont des polynômes d'interpolation de degré 3 que l'on raccorde en chaque points. Plus précisément, en supposant $x_1 < x_2 < \cdots < x_n$, on considère une suite de polynômes de degré 3 qui sur chaque intervalle $[x_i, x_{i+1}]$ telle que le polynôme passe en x_i et x_{i+1} et telle que les dérivées premières et secondes soient continues en ces points. Montrer que si l'on impose des conditions au bord (en x_1 et x_n) soit sur les dérivées premières, soit sur les dérivées secondes (par exemple qu'elles soient nulles), alors la suite de splines est unique.

iii. Demander que la fonction interpolatrice passe exactement par tous les points est une exigence contraignante qui finalement a peu de sens dans le cadre assez général de données entachées d'erreurs, voire la présence de données aberrantes. On préférera alors utiliser une méthode d'approximation avec la volonté que la fonction d'approximation, que l'on supposera au moins $\mathcal{C}^2$, n'ait pas de brusques variations. Un choix possible est de choisir une fonction f qui minimise la somme des carrés pénalisée :

$$S_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - f(x_i)\Big)^2 + \lambda \cdot \int_{-\infty}^{\infty}\left|\frac{\partial^2 f}{\partial x^2}(x)\right|^2 dx,$$

où $\lambda > 0$. On peut montrer que si l'on suppose que f appartient à l'espace des fonctions continûment dérivables et de dérivée seconde de carré intégrable (on admet que cet espace est un espace de Hilbert) alors il existe une unique fonction minimisant S_λ. On peut également montrer qu'une suite (appelée "smoothing"-splines) de polynômes de degré 3 entre chaque x_i et dont les dérivées secondes (donc également premières) sont continûment raccordés en x_i minimise $S_\lambda(.)$. La fonction formé par cette suite ne passe plus forcément par les points (x_i, y_i). Que se passe-t-il quand λ est proche de 0 ? Et lorsque λ tend vers l'infini ? Expliquer comment faire numériquement pour obtenir cette fonction.

Exercice 4.8

(***) [Erreur sur les règresseurs ou "total least squares"] On considère pour simplifier le modèle de règression linèaire simple

$$Y_i = \mu + \beta Z_i + \varepsilon_i, \quad i = 1, \cdots, n$$

avec les hypothèses habituelles. Mais on suppose que l'on observe les valeurs de la variable Z avec erreurs. En fait on observe seulement

$$\overline{Z}_i = Z_i + \eta_i$$

où les η_i sont des erreurs gaussiennes indèpendantes entre elles et indèpendantes des ε_i. On note σ_z^2 leur variance.

On peut ècrire bien-sûr

$$Y_i = \mu + \beta\overline{Z}_i + (\varepsilon_i - \beta\eta_i)$$

qui paraÃ®t être un modèle de règression linèaire à condition de poser $\overline{\sigma}^2 = \sigma^2 + \beta^2\sigma_z^2$.
1. Montrer que ce n'est pas le cas et que ce n'est pas le cas non plus pour le modèle conditionnel au $\overline{Z}_i$. (Attention ! dans le modèle de règression les Z_i ne doivent pas être alèatoires).

2. On suppose que le rapport σ^2/σ_z^2 est connu (sinon on ne sait pas faire grand chose). On suppose ègalement pour simplifier que $\mu = 0$. Montrer que sans perte de gènèralitè on peut supposer $\sigma^2 = \sigma_z^2$
3. Montrer que le modèle est alors un modèle statistique dont les paramètres sont $\beta, \sigma^2, Z_1, \ldots, Z_n$
4. Montrer que les estimateurs du maximum de vraisemblance des paramètres sont obtenus en minimisant

$$\sum_{i=1}^{n} dist^2(p_i, D(\beta))$$

où $dist$ est la distance euclidienne dans $\mathbb{R}^2$, p_i est le point de coordonnèes $(\overline{Z}_i, Y_i)$ et $D(\beta)$ est la droite de pente β.
5. Montrer que la solution de ce problème est obtenue par l'analyse en composante principale, ou SVD (single value decompositon) de la matrice

$$X = [Y \mid \overline{Z}]$$

Plus prècisèment, $(-1, \beta)'$ est le second vecteur propre de la matrice $X'X$.

Exercice 4.9

(**) [Test de Chow] Ceci est un test souvent utilisé en économétrie. On veut tester l'hypothèse H_0 : un modèle de type $Y = X\theta + \varepsilon$ régit les données, contre l'hypothèse H_1 : les données sont régies par 2 modèles : $Y = X\theta_1 + \varepsilon$ pour une partie des données et $Y = X\theta_2 + \varepsilon$ pour l'autre partie des données.
Une manière simple d'illustrer un tel test est l'exemple suivant : on dispose d'un nuage de points (i, Y_i) pour $i = 1, \cdots, n$, modélisant par exemple l'évolution temporelle d'une variable économique. On se pose la question de savoir s'il vaut mieux représenter ce nuage de points par une droite (hypothèse H_0) ou bien s'il existe une rupture dans les données qui impliquerait que ce nuage de points doive être modélisé par deux droites (hypothèse H_1).
Pour rendre plus simple l'exposition de ce test, on supposera donc que si X est la matrice de taille (n, p) et de rang p contenant les valeurs prises par les variables explicatives (dont le vecteur colonne $(1, 1, \cdots, 1)'$), il existe $p \leq n_1 \leq n - p$ connu tel que :

$$Y^{(1)} = X^{(1)}\theta_1 + \varepsilon^{(1)} \qquad \text{et} \qquad Y^{(2)} = X^{(2)}\theta_2 + \varepsilon^{(2)},$$

où $Y^{(1)} = (Y_i)_{1 \leq i \leq n_1}$ et $Y^{(2)} = (Y_i)_{n_1+1 \leq i \leq n}$, $\varepsilon^{(1)} = (\varepsilon_i)_{1 \leq i \leq n_1}$ et $\varepsilon^{(2)} = (\varepsilon_i)_{n_1+1 \leq i \leq n}$, $X^{(1)}$ et $X^{(2)}$ étant des matrices de tailles respectives (n_1, p) et $(n - n_1, p)$, supposées être de rang p, et qui sont les sous-matrices de X telles que $X' = [X^{(1)'}, X^{(2)'}]$. Le problème de test est alors

$$H_0 \ : \ \theta_1 = \theta_2 = \theta \qquad \text{contre} \qquad H_0 \ : \ \theta_1 \neq \theta_2.$$

La statistique de Chow est la statistique :

$$\widehat{C} := \frac{\frac{1}{p}\left(\|Y - X\widehat{\theta}\|_n - \|Y^{(1)} - X^{(1)}\widehat{\theta}_1\|_{n_1} - \|Y^{(2)} - X^{(2)}\widehat{\theta}_2\|_{n-n_1}\right)}{\frac{1}{n-2p}\left(\|Y^{(1)} - X^{(1)}\widehat{\theta}_1\|_{n_1} + \|Y^{(2)} - X^{(2)}\widehat{\theta}_2\|_{n-n_1}\right)},$$

où $\|\cdot\|_k$ désigne la norme euclidienne classique dans $\mathbb{R}^k$. Sous les postulats **P1-4**, déterminer la loi de $\widehat{C}$.

Chapitre 5

Problèmes spécifiques à l'analyse de la variance

Ce chapitre présente la notion d'interaction entre deux facteurs, les stratégies de test en analyse de la variance, les comparaisons multiples de moyennes et la notion de facteurs croisés et hiérarchisés.

1 Cadre général

Comme nous l'avons vu précédemment, l'analyse de la variance consiste à expliquer une variable quantitative Y par un certain nombre de variables qualitatives ou facteurs. Considérons deux exemples :

a. Comparaison variétale : la variable à expliquer est le rendement, elle peut être expliquée par

   - 2 facteurs : variété × lieu ;
   - 3 facteurs : variété × lieu × année ;
   - 5 facteurs : famille génétique × individu × lieu × année × testeur.

b. Etude du salaire annuel moyen d'un cadre. La variable à expliquer est le salaire annuel que l'on peut expliquer par

- 7 facteurs : domaine d'activité × tranche d'âge × taille de l'entreprise × région d'activité × niveau de diplôme × fonction exercée × sexe.

Comme nous le voyons ci-dessus, un nombre de facteurs élevé peut permettre de s'adapter finement à des situations relativement complexes et il ne faut pas hésiter à les utiliser. Les seules limitation sont 1/ de savoir manipuler proprement des modèles à plusieurs facteurs et 2/ de savoir également manipuler des jeux de données non-équirépétés. Dans l'exemple des 7 facteurs il est hautement improbable que le nombre d'individus de l'échantillon qui sont dans le domaine i, dans la tranche j, pour une taille k, un région l, un niveau m, une fonction n et un sexe o ne dépende pas de i, j, k, l, m, n, o. Ce sont donc ces deux difficultés qui vont être abordées dans ce chapitre.

Nous allons considérer d'abord le cas le plus classique de deux facteurs croisés. La raison pour laquelle ce modèle s'appelle ainsi ne sera expliquée qu'à la section 3.2

2 Deux facteurs croisés

2.1 Présentation

Plaçons-nous dans l'exemple historique des comparaisons de variétés agricoles (le problème sur lequel Ronald Fisher à débuté sa carrière). Pour comparer par exemple des variétés (en nombre I) de céréales selon un critère quantitatif comme le rendement, on dispose d'abord d'un certain nombre (J) de champs qui sont répartis dans diverses zones géographiques, par exemple : un champ en Brie, un en Beauce etc... Supposons maintenant que chaque champ est découpé en un grand nombre de parcelles et que la variété i ($i = 1, \cdots, I$) est expérimentée n_{ij} fois dans le lieu j ($j = 1, \cdots, J$). Cela signifie que l'on a semé n_{ij} parcelle avec cette variété et que ces parcelles ont été ensuite cultivées puis récoltées avec une mini-moissonneuse (on verra dans la partie consacrée au plans d'expériences que la réalité est souvent plus complexe). Mathématiquement on supposera que toutes les combinaisons sont représentées, c'est-à-dire que n_{ij} n'est jamais nul et même qu'il existe au moins une combinaison i, j qui est vue au moins deux fois.

Dans un premier temps, on pose un modèle général dépendant de la valeur du couple (i, j) :

$$Y_{ijk} = \theta_{ij} + \varepsilon_{ijk}, \quad \text{où} \tag{5.1}$$

- i est l'indice de variété variant de 1 à I ;

- j est l'indice de lieu variant de 1 à J ;
- k est l'indice de répétition variant de 1 à n_{ij} ;
- Y_{ijk} est le rendement mesuré dans la k-ième parcelle semée avec la variété i dans le lieu j.

Par exemple, supposons que $I = 2$, $J = 3$ et $n_{ij} = 2$ pour tout i, j. Le modèle s'écrit alors matriciellement :

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{13} \\ \theta_{21} \\ \theta_{22} \\ \theta_{23} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{131} \\ \varepsilon_{132} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{231} \\ \varepsilon_{232} \end{pmatrix}$$

Pour l'instant, le modèle que nous avons posé est en fait un modèle d'analyse de la variance à un facteur : le facteur produit variété×lieu. Ce facteur possède $I \times J$ modalités. Comme on l'a vu dans le chapitre les différentes valeurs θ_{ij} du paramètre θ sont estimées par les moyennes empiriques correspondantes, soit :

$$\widehat{\theta}_{ij} = Y_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \ .$$

Pour faire apparaître les deux facteurs originels, on définit différents effets par ce que nous appellerons la "décomposition marginale" (rappelons que, toute chose étant égales par ailleurs, l'on utilise la notation "." à la place d'un indice d'une variable pour signaler que l'on a effectué une moyenne sur toutes les valeurs possibles de cet indice) :

- la moyenne générale $\mu = \theta_{..}$, estimée par $\widehat{\theta}_{..}$;
- l'effet différentiel (en écart à la moyenne) de la modalité i du premier facteur $\alpha_i = \theta_{i.} - \theta_{..}$, estimé par $\widehat{\theta}_{i.} - \widehat{\theta}_{..}$;
- l'effet différentiel de la modalité j du deuxième facteur $\beta_j = \theta_{.j} - \theta_{..}$, estimé par

$\widehat{\theta}_{.j} - \widehat{\theta}_{..}$;

– la quantité manquante pour "arriver" à θ_{ij} est appelé *l'interaction* . En effet, il n'y a pas de raison que l'on ait $\theta_{ij} = \theta_{i.} + \theta_{.j} + \theta_{..}$. On rajoutera donc ce terme d'interaction γ_{ij} tel que :

$$\gamma_{ij} = \theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..} = (\theta_{ij} - \theta_{..}) - (\theta_{i.} - \theta_{..}) - (\theta_{.j} - \theta_{..})$$

Au final, le modèle initial (5.1) : $Y_{ijk} = \theta_{ij} + \varepsilon_{ijk}$ se réécrit sous la forme :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{5.2}$$

avec $k \in \{1, \cdots, n_{ij}\}$ pour $(i,j) \in \{1, \cdots, I\} \times \{1, \cdots, J\}$, et les contraintes suivantes que, nous avons implicitement défini :

i. $\sum_{i}^{I} \alpha_i = 0$ et $\sum_{j}^{J} \beta_j = 0$;

ii. pour tout $j = 1, \cdots, J$, on a $\sum_{i} \gamma_{ij} = 0$;

iii. pour tout $i = 1, \cdots, I$, on a $\sum_{j} \gamma_{ij} = 0$.

On peut également écrire ce modèle matriciellement dans le cas de l'exemple précédent, à deux facteurs, avec $I = 2$, $J = 3$ et $n_{ij} = 2$ pour tout i, j, soit :

$$\begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{131} \\ Y_{132} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{23} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{131} \\ \varepsilon_{132} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{231} \\ \varepsilon_{232} \end{pmatrix}.$$

La matrice X ici n'est pas de rang plein, car l'on n'a pas tenu compte de toutes les relations entre les différents paramètres. On verra au chapitre 7 l'étude détaillée de tel modèles dit non-réguliers, mais ici il est beaucoup plus simple de considérer le modèle

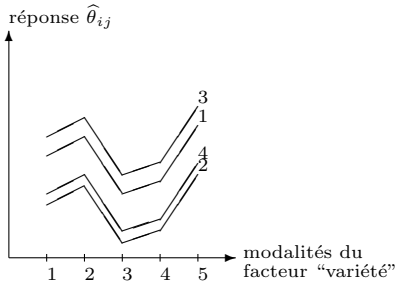
(5.2) comme une réécriture du modèle (5.1) qui, lui, est régulier et ne pose pas de problèmes particuliers. Le modèle (5.2) prend une nature très différente suivant que la partie γ est présente ou non. C'est l'objet de la définition suivante.

Définition 5.1 *Soit le modèle (5.2) avec les contraintes ci-dessus. Lorsque les paramètres d'interaction γ_{ij} sont nuls pour tout $(i,j) \in \{1,\cdots,I\} \times \{1,\cdots,J\}$, le modèle est dit additif. Sinon, on dit que le modèle est avec interaction, ou encore on parle de "modèle général". Enfin, on parle d'"effets principaux" pour tout ce qui est relatif aux paramètres α_i et β_j.*

Il y a une grande différence entre les deux types de modèles : général et additif . En premier lieu, la dimension paramétrique (le nombre de paramètres distincts à estimer) est beaucoup plus faible dans le cas du modèle additif. En effet, elle vaut, dans le cas du modèle additif, $I + J - 1$ (par exemple si $I = J$=10, dim= 19) et dans le cas du modèle général $1 + (I-1) + (J-1) + (I-1)(J-1) = I \times J$ (exemple, dim= 100). Ce premier élément nous indique un premier avantage substantiel à chercher à se ramener au modèle additif.

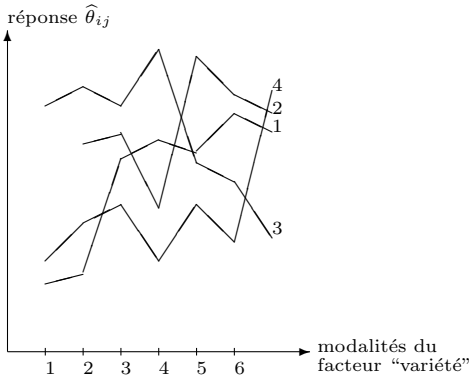
En second lieu, ces deux types de modèles correspondent à des comportements différents entre les facteurs. On peut ainsi représenter les différentes valeurs de $\widehat{\theta}_{ij}$ du modèle en fonction des modalités i et j des deux facteurs. Illustrons ceci par un exemple dans lequel un facteur (par exemple la variété) a 6 modalités et l'autre (par exemple son lieu de culture) en a 4 :

Modèle additif



Les courbes sont à peu près parallèles. Cela signifie que pour toutes les modalités de l'un des facteurs, la différence de réponse moyenne $\widehat{\theta}_{ij}$ entre deux modalités fixées de l'autre facteur reste constante. Par exemple, quelque soit la variété de la plante, la différence de réponse entre les lieux 3 et 1 est à peu près constante. Plus généralement, on peut également écrire que pout tout i donné, on a pour tout j, $\widehat{\theta}_{ij} - \widehat{\theta}_{i1} = \beta_j - \beta_1$ la réponse $\widehat{\theta}_{ij}$ est bien l'addition des effets de i et des effets de j : le modèle est additif.

Modèle général avec interaction



Le comportement des différentes courbes ne semble plus présenter de particularité. La réponse $\widehat{\theta}_{ij}$ semble dépendre de chaque composant i et j, les propriétés d'additivité précédentes ne sont plus applicables.

2.2 Modèle avec interaction dans le cas équirépété

On va supposer que chaque combinaison (i, j) des différentes modalités des deux facteurs est expérimentée le même nombre de fois K : c'est en ce sens que l'on appellera le modèle "équirépété". Pour des question de dimension on supposera de plus que $K > 1$ et on utilisera le modèle (5.2) dans toute sa généralité :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \ , \ i = 1, \cdots, I \ , \ j = 1, \cdots, J \ , \ k = 1, \cdots, K,$$

avec les contraintes décrites plus haut. On veut tester les différentes hypothèses stipulant la présence ou non d'un des effets principaux ou de l'interaction. Plus précisément on définit les hypothèses :

- $H_0^{(1)}$: "tous les coefficients α_i sont nuls" : test de l'effet principal du premier facteur ;
- $H_0^{(2)}$: "tous les coefficients β_i sont nuls" : test de l'effet principal du second facteur ;
- $H_0^{(3)}$: "tous les coefficients γ_{ij} sont nuls" : test de l'interaction.

Notons qu'en fait toutes ces hypothèses peuvent s'exprimer sans problème dans le modèle (5.1). On peut donc appliquer les résultats du chapitre 3.

Une des principales difficultés est de choisir une bonne représentation des vecteurs qui vont intervenir dans nos calculs. En effet, l'espace $E := \mathbb{R}^n$ des observations ($n = I \cdot J \cdot K$) est en fait constitué de tableaux à trois indices i, j, k. Dans un premier temps, nous conserverons cette structure qui rend la représentation plus claire. Nous écrirons

$$E = [\{Y_{ijk} \ : \ i = 1, \cdots, I; \ j = 1, \cdots, J; \ k = 1, \cdots, K\}]$$

et nous noterons en abrégé $(Y_{ijk})_{ijk}$ un élément de E. En effet, comme nous l'avons déjà mentionné au chapitre 1, le terme "vecteur" veut dire élément d'un espace vectoriel : espace stable par addition et multiplication scalaire. Il n'est pas obligatoire que les données soient sous forme "vecteur colonne". Ce n'est que lorsque nous écrirons des formules de calcul matriciel que nous serons obligé d'utiliser ses conventions. Les vecteurs de E seront alors "déroulés" en ordre lexicographique. On écrira alors

$$\big((Y_{ijk})_{ijk}\big)' = \big(Y_{111}, \cdots, Y_{11K}, Y_{121}, \cdots, Y_{12K}, Y_{131}, \cdots, Y_{1JK}, Y_{211}, \cdots, Y_{IJK}\big)' .$$

Nous utilisons la norme euclidienne standard sur E et nous définissons les sous-espaces vectoriels de E suivants :

- $E_0 = [\mathbb{1}] = \{(Y_{ijk})_{ijk} \in E :$ il existe m un réel tel que $Y_{ijk} = m$ pour tout $i, j, k\}$.
  C'est le sous-espace vectoriel de $\mathbb{R}^n$ formé des vecteurs dont toutes les coordonnées sont égales ;

- $E_1 = \{(Y_{ijk})_{ijk} \in E :$ il existe $a_1, \cdots, a_I$ avec $\sum_i a_i = 0$ tel que $Y_{ijk} = a_i$ pour tout $i, j, k\}$.
  E_1 est le sous-espace vectoriel de E formé des vecteurs dont les coordonnées ne dépendent que de i et dont la somme est nulle ;

- $E_2 = \{(Y_{ijk})_{ijk} \in E :$ il existe $b_1, \cdots, b_J$ avec $\sum_j b_j = 0; Y_{ijk} = b_j$ pour tout $i, j, k\}$.
  E_2 est le sous-espace vectoriel de E formé des vecteurs dont les coordonnées ne dépendent que de j et dont la somme est nulle ;

- $E_3 = \{(Y_{ijk})_{ijk} \in E :$ il existe $c_{11}, \cdots, c_{IJ}$ avec $\forall i, \ \sum_j c_{ij} = 0; \forall j, \ \sum_i c_{ij} = 0$ et $Y_{ijk} = c_{ij}$ pour tout $i, j, k\}$.
  E_3 est le sous-espace vectoriel de E formé des vecteurs dont les coordonnées ne dépendent que de i et de j et dont la somme est nulle suivant i ou suivant j ;

On a alors les propriétés suivantes qui sont faciles à montrer :

- E_0, E_1 E_2 et E_3 sont des sous-espaces orthogonaux entre eux ;
- $E_0 + E_1$ (somme des sous-espaces) correspond au modèle à un seul facteur : le premier ;
- $E_0 + E_2$ correspond au modèle restreint au second facteur ;
- $E_0 + E_1 + E_2$ correspond au modèle additif ;
- $E_0 + E_1 + E_2 + E_3$ correspond au modèle complet

("correspond" veut dire que l'espace considéré est l'espace dans lequel réside l'espérance de la réponse Y dans le modèle correspondant).

L'orthogonalité entre ces sous-espaces implique que les projecteurs se somment, dans le sens où, par exemple,

$$P_{E_0+E_1+E_2+E_3}(Y) = P_{E_0}(Y) + P_{E_1}(Y) + P_{E_2}(Y) + P_{E_3}(Y).$$

En utilisant le fait que dans un modèle d'analyse de la variance à un facteur, les estimateurs sont des simples moyennes, on obtient :

- $P_{E_0}(Y) = \left(Y_{...}\right)_{ijk}$,
- $P_{E_0+E_1}(Y) = \left(Y_{i..}\right)_{ijk}$,
- $P_{E_0+E_2}(Y) = \left(Y_{.j.}\right)_{ijk}$,
- $P_{E_0+E_1+E_2+E_3}(Y) = \left(Y_{ij.}\right)_{ijk}$,

Par combinaison linéaire il est facile d'en déduire les expressions de chacun des projecteurs.

Considérons maintenant par exemple le test de la nullité de l'interaction, c'est-à-dire le test de l'hypothèse $H_0^{(3)}$. Au dénominateur de la statistique $\widehat{F}$ du test de Fisher associé, on trouve, comme toujours le carré moyen résiduel qui vaut ici

$$CMR = \frac{1}{n - I.J}\|P_{(E_0+E_1+E_2+E_3)^\perp}(Y)\|^2 = \frac{1}{n - I.J}\sum_{ijk}(Y_{ijk} - Y_{ij.})^2.$$

Au numérateur, avant division par les degrés de liberté $(I-1)(J-1)$, on trouve la différence entre la somme des carrés du modèle additif :

$$\|P_{(E_0+E_1+E_2)^\perp}(Y)\|^2,$$

et la somme des carrés précédente

$$\|P_{(E_0+E_1+E_2+E3)^\perp}(Y)\|^2.$$

Une application du Théorème de Pythagore donne donc pour le numérateur N,

$$N = \frac{1}{(I-1)(J-1)}\|P_{E3}(Y)\|^2 = \frac{1}{(I-1)(J-1)}\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2$$

et donc

$$\widehat{F} = \frac{(n-I.J)\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2}{(I-1)(J-1)\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2}.$$

En considérant le cas des effets principaux on obtient, par des démonstrations analogues, la *table d'analyse de la variance* suivante, qui donne tous les éléments de construction des différents tests.

Source	Somme de carrés	Degrés de liberté	$\widehat{F}$
Facteur 1	$\sum_{i,j,k}(Y_{i..} - Y_{...})^2$	$I-1$	$\frac{(n-I.J)}{(I-1)}\frac{\sum_{i,j,k}(Y_{i..} - Y_{...})^2}{\sum_{i,j,k}(Y_{i,j,k} - Y_{ij.})^2}$
Facteur 2	$\sum_{i,j,k}(Y_{.j.} - Y_{...})^2$	$J-1$	$\frac{(n-I.J)}{(J-1)}\frac{\sum_{i,j,k}(Y_{.j.} - Y_{...})^2}{\sum_{i,j,k}(Y_{i,j,k} - Y_{ij.})^2}$
Fac 1 × Fac 2	$\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2$	$(I-1)(J-1)$	$\frac{(n-I.J)\sum_{i,j,k}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2}{(I-1)(J-1)\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2}$
Résidu.	$\sum_{i,j,k}(Y_{ijk} - Y_{ij.})^2$	$n - I \cdot J$	

(pour ne pas alourdir ce tableau, nous n'avons pas complété les carrés moyens qui s'obtiennent naturellement en divisant la somme des carrés par les degrés de liberté). Il est clair que la statistique $\widehat{F}$ présentée pour le Facteur 1 est utilisée pour tester l'hypothèse $H_0^{(1)}$, et de même avec Facteur 2 et Facteur 1 × Facteur 2 pour les hypothèses $H_0^{(2)}$ et $H_0^{(3)}$.

Remarque 1 : Dans la table d'analyse de la variance ci-dessus, toutes les sommes sont exprimées en fonction des trois indices i, j, k alors que parfois l'expression dans la somme n'en depend pas. Ce n'est pas un erreur de notre part, mais un choix délibéré de la forme la plus simple à mémoriser. En effet sous la forme ci-dessus, les sommes

apparaissent comme des carrés de norme euclidienne et il n'y a aucun coefficient en facteur. Un autre choix aurait été d'écrire par exemple

$$SC_1 = \sum_{i,j,k} (Y_{i..} - Y_{...})^2 = J \cdot K \cdot \sum_i (Y_{i..} - Y_{...})^2$$

mais les formules sont plus difficiles à mémoriser sous cette forme.

Remarque 2 : La table d'analyse de la variance a-t-elle un intérêt ? Dans le tableau ci-dessus, seules les valeurs des $\widehat{F}$ (et les degrés de liberté dans une moindre mesure pour comprendre ce qui se passe) ont vraiment un intérêt, et encore, souvent, on se limitera à lire le niveau de signification (P-value). Pourtant ce tableau est traditionnel et il est donné de manière systématique par tous les logiciels de statistique. On ne peut donc faire l'économie de sa présentation. L'opinion personnelle des auteurs est que l'origine de cette tradition vient du temps où ces calculs étaient effectués avec la machine à calculer à manivelle : sur ces machines dont la plus utilisée en France était de marque "Vaucanson", il fallait quelque secondes pour une addition, plus d'une minute pour une multiplication. La puissance d'un telle machine était de l'ordre de 0.01 flops (floating operation per second). Comme votre ordinateur tutoie sans aucun doute les giga-flops, les échelles ont changé... À cette époque où il fallait une journée pour faire les calculs d'une analyse de la variance, il était primordial de présenter les calculs intermédiaires dans un tableau. Avec le temps, les statisticiens ont pris l'habitude de se référer à ce tableau qui a un intérêt descriptif et l'usage est resté.

2.3 Modèle additif à deux facteurs dans le cas équirépété

Le cas du modèle additif se traite avec les mêmes techniques. Les différences principales sont les suivantes :

- le nombre de répétitions K de chaque combinaison peut être égal à 1. Dans ce cas particulier d'ailleurs le modèle additif est imposé car le modèle complet serait saturé ;
- on ne définit pas E_3 et $H_0^{(3)}$;
- on montre que :

$$\widehat{Y} = P_{E_0+E_1+E_2}(Y) = P_{E_0}(Y) + P_{E_1}(Y) + P_{E_2}(Y).$$

Tout ceci permet d'en déduire la table d'analyse de la variance suivante :

Source	Somme de carrés	Degrés de liberté	$\widehat{F}$
Facteur 1	$\sum_{i,j,k}(Y_{i..} - Y_{...})^2$	$I-1$	$\frac{(n-I-J+1)\sum_{i,j,k}(Y_{i..} - Y_{...})^2}{(I-1)\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2}$
Facteur 2	$\sum_{i,j,k}(Y_{.j.} - Y_{...})^2$	$J-1$	$\frac{(n-I-J+1)\sum_{i,j,k}(Y_{.j.} - Y_{...})^2}{(J-1)\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2}$
Résiduelle	$\sum_{i,j,k}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2$	$n-I-J+1$	

2.4 Quel modèle choisir ?

Les tests à faire en analyse de la variance ne sont pas uniques et plusieurs écoles existent. En premier lieu il faut regarder si l'interaction est significative car il est beaucoup plus intéressant d'utiliser un modèle additif pour décrire les données. **Notre attitude par rapport à l'interaction sera tout ou rien : ou elle est présente, ou elle est totalement absente.** Il existe des modèles intermédiaires dits de structuration de l'interaction qui permettent d'éviter un choix si violent, nous renvoyons à Denis [22]. A la suite de ce premier test, on considère les deux situations suivantes :

Cas 1 : L'interaction est significative, c'est-à-dire que l'hypothèse $H_0^{(3)}$ est rejetée par le fait que le $\widehat{F}$ est trop élevé. Alors il est clair que les deux facteurs sont pertinents et qu'aucun des deux ne peut être enlevé du modèle. Cependant, votre logiciel préféré (quel qu'il soit) vous proposera toujours le test sur les effets principaux (c'est-à-dire sur les hypothèses $H_0^{(1)}$ et $H_0^{(2)}$). Ce test n'est en aucun cas le test de l'absence de l'effet puisqu'il est déjà présent dans l'interaction, mais beaucoup d'utilisateurs s'accordent pour dire que ce test a un intérêt descriptif : il permet d'apprécier par l'intermédiaire des $\widehat{F}$, le rapport des ordres de grandeur entre les effets principaux et l'interaction. Cette démarche est à rapprocher de l'estimation de composantes de la variance dans des modèles mixtes (voir chapitre 10).

Cas 2 : L'interaction est non significative (donc $H_0^{(3)}$ est acceptée). Dans ce cas, on doit encore tester les effets principaux. La logique voudrait que l'on se place alors dans le modèle additif. Cependant, particulièrement quand le jeu de données est important, cela est déconseillé car on désire se protéger contre un possible manque de puissance du test de l'interaction. On ne désire pas qu'une faible interaction non décelée vienne fausser l'estimateur $\widehat{\sigma}^2$ de σ^2 comme ce serait le cas dans le modèle additif. Pour ces raisons, on garde l'estimateur $\widehat{\sigma}^2$ du modèle complet et on définit l'absence d'effets principaux comme la nullité des effets α_i et β_j définis dans la décomposition marginale (donc en reprenant les statistiques associées aux tests de $H_0^{(1)}$ et $H_0^{(2)}$).

2.5 Différences entre des expériences équirépétées et non-équirépétées

Jusqu'à présent, nous avons cherché délibérément à masquer les difficultés qui apparaissent dans le cas non-équirépété. Pourtant le cas non-équirépété reste souvent le cas général. En effet il existe de nombreuses situations où l'on prévoit de réaliser une expérience équirépétée, mais au moment de la réalisation pour diverses raisons certaines des combinaisons donnent lieu a un résultat divergent. On doit les éliminer : on parle de *données manquantes* et le bel équilibre du départ est rompu. Voyons les propriétés qui restent dans ce cas.

– Quand les données sont équirépétées, il y a unicité de la définition des effets principaux et il y a unicité de la table d'analyse de la variance. De plus, les estimateurs sont des moyennes ordinaires, et on a en particulier :

$$\widehat{\theta_{i.}} = Y_{i..} \quad \text{et donc} \quad \alpha_i = Y_{i..} - Y_{...}$$

Ce genre de calcul peut être mené par des programmes qui sont adaptés à l'équirépétition, mais qui ne tournent pas dans le cas non équirépété.

– Quand les données ne sont plus équirépétées, il n'y a plus unicité de la définition des effets principaux. La définition que nous avons donnée en section 2.1 correspond à un choix : **c'est la décomposition dite de type III dans la terminologie de SAS qui s'est peu a peu imposée**. Nous avons retenu la solution qui nous paraît être celle adoptée par la majorité des statisticiens et qui est clairement adaptée au cas ou la non-équirépétition est le fruit du hasard à travers l'existence de données manquantes. On consultera Azaïs [5] pour la définition des autre types de décomposition. La décomposition de type I est abordée dans l'exercice 5. Dans tous les cas, la table d'analyse de la variance n'a en aucun cas l'expression simple du cas équirépété. Plus précisément, dans le cas non-équirépété, la somme des carrés associée à l'interaction garde tout de même une définition unique : elle est définie comme la différence de sommes de carrés résiduelles entre modèle additif et modèle complet. Cependant les estimateurs du modèle additif n'ont pas des expressions simples sous forme de moyennes ou bien de moyennes de moyennes. Il faut alors résoudre un système d'équations linéaires. Cette somme a donc une expression compliquée. La somme des carrés associée à un des deux effets principaux découle du choix que nous avons fait de la définition de cet effet. Elle correspond par exemple pour le premier facteur au test de l'hypothèse

$$\text{"}\theta_{i.} \text{ ne dépend pas de } i\text{"}$$

Sa forme est encore complexe. Le lecteur curieux peut consulter le livre de Searle [54] p. 87-93 : l'expression de la somme des carrés associée au premier facteur

dans la table d'analyse de la variance est (par exemple) :

$$SC = \sum_{i=1}^{I} w_i \left(\left[\frac{\sum w_i \widehat{\theta}_{i.}}{\sum w_i} \right] \widehat{\theta}_{i.} \right)^2$$

où, pour chaque $i = 1, \cdots, I$, le poids w_i est défini par le fait que :

$$\text{Var}\left(\widehat{\theta}_{i.}\right) = \frac{\sigma^2}{w_i} \text{ et donc } \frac{\sigma^2}{w_i} = \frac{\sigma^2}{J^2} \sum_{j=1}^{I} \frac{1}{n_{ij}}.$$

Notons que les estimateurs dans le modèle complet sont construits à partir de moyennes de moyennes. On construit les $\widehat{\theta}_{ij} = Y_{ij.}$ qui sont des moyennes et on fait ensuite des moyennes ($\widehat{\theta}_{i.}$ par exemple) de ces moyennes. Remarquez bien qu'une moyenne de moyenne n'est pas en général une moyenne.

3 Extensions

La partie précédente donne les grands traits pré-requis à toute analyse de la variance avec deux facteurs. Dans cette partie nous traitons certaines extensions du modèle croisé à deux facteurs.

3.1 Comparaisons multiples

Si on teste la significativité d'un facteur (appelons le "traitement" par exemple), deux résultats peuvent se produire :

- Si l'effet (facteur) "traitement" est non significatif, on n'ira pas plus loin dans l'analyse et on considérera l'expérience comme négative pour ce qui concerne l'influence de ce facteur.

- Si l'effet traitement est significatif, on désire en général pousser l'analyse plus loin en classant les différents traitements ou en les comparant à un témoin.

Si on est en présence de I traitements différents (par exemple $I = 10$), il faut examiner les $I(I-1)/2$ (par exemple $= 45$) comparaisons possibles de deux traitements deux à deux. Pour comparer deux traitements définis a priori, par exemple les modalités 1 et 2, on peut utiliser le test de Student T général défini lors du chapitre 3. Dans le modèle (5.2), on désire tester l'hypothèse $\alpha_1 = \alpha_2$ qui correspond en fait à l'hypothèse $\theta_{1.} = \theta_{2.}$ dans le modèle (5.1). On peut donc utiliser le test de Fisher correspondant

ou, ce qui revient au même, poser

$$\widehat{T}_{12} = \frac{\widehat{\theta_{1.}} - \widehat{\theta_{2.}}}{\sqrt{\widehat{\mathrm{Var}}\,(\widehat{\theta_{1.}} - \widehat{\theta_{2.}})}}.$$

On sait que sous l'hypothèse d'égalité des deux traitements

$$\widehat{T}_{12} \overset{\mathcal{L}}{\sim} T(n - IJ).$$

Ce test de Student est parfaitement valide pour comparer deux traitements choisis a priori. En revanche, il n'est plus du tout utilisable pour comparer le traitement qui donne en apparence les résultats les meilleurs, avec celui qui donne en apparence les résultats les plus mauvais. En effet, cela revient à comparer tous les traitements entre eux et chaque test a une probabilité α (le niveau du test) de déclarer présente une différence qui n'existe pas. Au total, sur les $I(I-1)/2$ comparaisons, la probabilité d'en déclarer une significative par "hasard" devient importante. Pour contrôler un risque global sur les $I(I-1)/2$ comparaisons deux à deux ou sur les $(I-1)$ comparaisons à un témoin, il existe diverses méthodes. Commençons par les méthodes de comparaison deux à deux.

Méthodes de comparaison deux à deux :

i. Méthode de Tukey : elle est adaptée au cas équirépété. Elle fournit des intervalles de confiance simultanés pour les différences entre paramètres $\alpha_i - \alpha_j$ où $1 \leq i < j \leq I$ (le risque est global sur les $I(I-1)/2$ comparaisons). Dans le cas équirépété, c'est la méthode la plus précise.

ii. Méthode de Newman et Keuls : elle présente une légère modification de la méthode précédente, mais elle ne fournit plus d'intervalles de confiance.

iii. Méthode de Bonferroni : cette méthode est la plus simple et peut être appliquée dans tous les cas. Si on à $I(I-1)/2$ comparaisons à faire et que l'on veut un risque global de niveau α, on fait toutes les comparaisons par un test de Student classique, mais au niveau

$$\alpha' = \frac{\alpha}{I(I-1)/2}.$$

Cette méthode est particulièrement adaptée au cas où I est petit et le dispositif déséquilibré.

iv. Méthode de Scheffé : c'est une méthode très sûre qui consiste à construire un ellipsoïde de confiance pour le vecteur des paramètres θ ou pour une sous-partie

de ce vecteur (par exemple les comparaisons entre traitements). Cet ellipsoïde de confiance se projette en intervalles de confiance pour les différences. Là encore, on laissera à l'ordinateur le soin d'effectuer les calculs. La méthode de Scheffé a l'avantage (qui se paye par des intervalles légèrement plus grands) de donner des intervalles de confiance pour n'importe quelle combinaison des traitements.

Prenons l'exemple de trois traitements dans un dispositif équirépété que l'on veut tester avec un niveau α :

i. La méthode de Tukey donne des intervalles de confiance pour : $\alpha_1 - \alpha_2, \alpha_3 - \alpha_2, \alpha_1 - \alpha_3$ avec un risque global de α ;

ii. La méthode de Bonferroni donne un intervalle de confiance pour $\alpha_1 - \alpha_2$ de niveau $\alpha/3$; un intervalle pour $\alpha_2 - \alpha_3$ de niveau $\alpha/3$ et un intervalle pour $\alpha_3 - \alpha_1$ de niveau $\alpha/3$;

iii. La méthode de Scheffé donne des intervalles de confiance simultanés de risque global de α pour toutes les combinaisons linéaires possibles entre les paramètres. Par exemple, en plus des trois ci-dessus, on peut obtenir un intervalle de confiance pour : $\alpha_1 - 2\alpha_2 + \alpha_3$.

Méthodes de comparaison à un témoin :

i. Méthode de Bonferroni. C'est le même principe que précédemment, mais il y a maintenant $I - 1$ comparaisons à effectuer, et on prend donc pour niveau de confiance : $\alpha' = \alpha/(I - 1)$.

ii. L'équivalent de la méthode de Tukey dans le cas équilibré est la méthode de Dunnet, qui construit des intervalles de confiance pour les combinaisons

$$\alpha_i - \alpha_1 \text{ avec } 2 \leq i \leq t,$$

si on suppose que le facteur témoin est le traitement 1.

Tous les détails sur ces différentes méthodes peuvent se trouver dans l'ouvrage de Miller [46].

3.2 Plusieurs facteurs, facteurs croisés et hiérarchisés

Dans le cas où l'analyse porte sur plus de deux facteurs, on décompose la réponse en : effets principaux, interactions doubles, triples, etc ... et on utilise la même

stratégie que précédemment. On trouvera une définition mathématique de l'interaction entre un nombre quelconque de facteurs au chapitre 13. Un cas particulier doit toutefois être précisé : il faut distinguer le cas de facteurs croisés et de facteurs hiérarchisés :

Définition 5.2 *Deux facteurs sont dit croisés si chacun d'eux a un sens indépendamment de l'autre.*
Le facteur B est hiérarchisé au facteur A si un indice du facteur B ne signifie rien de concret, tant que l'on ne connaît pas l'indice associé du facteur A.

Pour bien distinguer entre ces deux types d'interaction, voici d'abord des exemples typiques de facteurs généralement croisés :

- variété $*$ lieu $\Longrightarrow$ pour prédire un rendement ;
- type de voiture $*$ type de trajet $\Longrightarrow$ pour prédire une consommation ;
- concentration $*$ température $\Longrightarrow$ pour prédire le rendement d'une réaction chimique.

Voici maintenant des exemples typiques de facteur hiérarchisés :

- produit/échantillon dans produit $\Longrightarrow$ pour un test bactériologique ;
- bloc/sous-bloc $\Longrightarrow$ en expérimentation variétale ;
- lapine/numéro de la portée /numéro du lapin dans la portée $\Longrightarrow$ en expérimentation animale.

Le notion de hiérarchie se détecte par le fait qu'il ne peut pas y avoir d'effet propre du numéro de lapin dans la portée, car il n'y a aucun rapport entre tous les lapins classés n^o4 dans les différentes portées. De même, il n'y a aucun rapport entre le sous-bloc n^o2 du premier bloc et le sous-bloc n^o2 du second bloc.

Quand on est en présence de facteurs hiérarchisés, il faut prendre soin de le déclarer avec la syntaxe propre au logiciel que l'on utilise. En effet, il ne faut pas dans ce cas, définir un effet propre du facteur hiérarchisé. La décomposition de la réponse θ_{ij} doit être différente de celle du modèle croisé vue au paragraphe 2.1 : ce qui serait l'effet principal du facteur hiérarchisé dit incorporé à l'interaction. On élimine alors dans le modèle les termes correspondants aux effets principaux seconds dans l'ordre de la

hiérarchie, pour obtenir le modèle suivant :

$$Y_{ijk} = \mu + \alpha_i + \gamma'_{ij} + \varepsilon_{ijk}, \tag{5.3}$$

avec les contraintes $\sum_i \alpha_i = 0$ et pour tout i, $\sum_j \gamma'_{ij} = 0$.

3.3 Tester l'inhomogénéité des variances

Les graphes :

- résidus ($\widehat{\varepsilon}_{ij}$) contre le niveau (ici i) de l'unique facteur ;
- résidu ($\widehat{\varepsilon}_{ijk}$) contre réponse estimée ($\widehat{Y_{ijk}}$) (s'il y a plusieurs facteurs) ;

peuvent montrer une plus grande dispersion dans certaines régions de l'expérience ; on suspectera alors la non validité du second postulat, c'est-à-dire une inhomogénéité des variances. Le test classique utilisé dans ce cas-là est le test de Bartlett basé sur une méthode de maximum de vraisemblance. Cependant ce test est déconseillé car il n'est pas robuste à la non normalité. On utilisera donc plutôt le test de Levene (voir [40]) ou sa modification basée sur les carrés. Dans le cas d'un unique facteur, le principe du test est le suivant :

Soit Y_{ij} la j-ème observation de la modalité i. On déduit de l'analyse de la variance les résidus $\widehat{\varepsilon}_{ij}$. On effectue alors une analyse de la variance sur les $|\widehat{\varepsilon}_{ij}|$. Si les variances sont homogènes, ces quantités doivent être d'espérance $(cte)\sigma$, c'est-à-dire constantes. Sinon, leur valeur "moyenne" varie avec la valeur de i. C'est donc la valeur du test de Fisher appliqué aux valeurs absolues des résidus qui donne le test de l'homoscédasticité. On peut également préférer travailler sur la variable $(\widehat{\varepsilon}_{ij})^2$ et dans le cas de plusieurs facteurs, on testera les différents effets principaux.

Comme en régression, les deux seuls remèdes en cas d'inhomogénéité des variances sont :

- les transformations de la variable Y avec les mêmes règles ;
- le recours au modèle linéaire généralisé.

3.4 Plans à mesures répétées

On définit dans beaucoup d'expériences la notion d'unité statistique. Cette notion, qui sera reprise en détail lors des chapitres consacrés aux plans d'expériences, est une notion subjective qui correspond à une unité de recueil de données. Dans les cas simples, l'unité correspond au recueil **d'une seule donnée**. L'unité est alors :

un groupe de femmes de même âge ; un arbre ; etc... Dans certaines situations plus complexes, on sait bien ce qui est dit jouer le rôle d'unité, par exemple, un individu (voir exemple informatique du chapitre 6), mais on va faire **plusieurs mesures** sur chaque unité (on va par exemple mesurer la taille à différents âges). Nous donnons ci-dessous deux exemples de ce type de situations.

Exemple 1 : Test bactériologique sur des fragments de dents (Calas 1993)

On soumet des fragments de dents à une contamination microbienne, puis à une désinfection à l'aide de différents produits (facteur "traitement"). Pour mesurer l'infection résiduelle, on observe au microscope électronique un certain nombre de régions ou "spots" du fragment de dent dont on compte le nombre de germes. Analysons les sources de variabilité de cette expérience :

- les fragments de dents sont différents les uns des autres. C'est la première source de variabilité ;
- le choix du spot dans la dent est aléatoire : on peut "tomber par hasard" sur une zone infectée ou non.

Du fait qu'il peut y avoir plusieurs mesures sur chaque fragment de dent, nous sommes donc en présence de mesures répétées. On dira de manière équivalente que l'on est en présence d'un plan à deux sources d'erreur. Ici la solution sera simple puisque le facteur d'intérêt (le traitement) ne varie pas quand varie l'indice de répétition (les différents spots). On a donc la possibilité de se ramener à un modèle classique en calculant les moyennes par fragment : l'indice de répétition a alors disparu et le modèle est un modèle linéaire ordinaire.

Exemple 2 : Moelleux de gâteaux (voir Cochran et Cox [18])

On désire comparer trois recettes de gâteau au chocolat quant à leur moelleux. Chaque pâte de gâteau est divisée en six parties qui sont cuites à six températures différentes. Le plan est alors de manière irréductible un plan avec deux sources d'erreur :

- l'erreur de composition de chaque pâte : erreur de pesée et de variabilité de composition des ingrédients (lait, oeufs, farine). Cette erreur est identique sur les 6 gâteaux issus de la même pâte.
- l'erreur de mesure.

Il faut alors utiliser un modèle mixte comme cela sera fait au chapitre 10.

4 Exemples traités par logiciels informatiques

4.1 Analyse de la variance à deux facteurs équirépétée

Les données que nous allons utiliser sont extraites de l'article de Calas *et al.* (1998) [17]. Dans cette expérience, on compare l'action de deux traitements (facteur `trait`) désinfectants sur des échantillons de racines de dents de vaches contaminées au préalable par deux sources de germes (facteur `germe`). La réponse est le nombre moyen de germes restants. Elle est mesurée par microscopie électronique. Pour des raisons d'homogénéité de la variance, on travaillera plutôt sur le logarithme de ce nombre (variable `LNBAC`). En fait, d'autres facteurs devraient être incorporés dans le modèle d'analyse : l'âge de la dent, la vache dont est issue la dent, etc..., mais par souci de simplicité nous les omettrons ici.

Logiciel Splus :

Traitons maintenant cet exemple avec Splus. Les commandes pour cette analyse de la variance peuvent être les suivantes :

```
attach(dents)
germe<-as.factor(germe)
trait<-as.factor(trait)
menuAov(LNBAC ~ germe*trait,plotResidVsFit.p=T,means.p=T)
graphsheet()
par(mfrow=c(1,2))
interaction.plot(germe,trait,response=LNBAC)
interaction.plot(trait,germe,response=LNBAC)
```

Commentaires sur ces commandes :

- les trois premières commandes permettent d'intégrer à la session ouverte le fichier `dents` (ce qui permettra d'appeler une variable du fichier directement `germe` et non `dents$germe`), et de bien préciser que les variables `germe` et `trait` sont des facteurs (qualitatifs).

- la quatrième commande réalise l'analyse de la variance de la variable quantitative `LNBAC` par les facteurs `germe` et `trait` en considérant qu'il y a un terme d'interaction. On a demandé, en plus de l'analyse de la variance "brute", d'afficher le graphe des résidus en fonction des valeurs prédites ainsi que les moyennes pour les différentes modalités de chacun des facteurs et de leurs interactions

(commande `means`, adapté au cas présent de l'équirépartition ; sinon, on aurait utiliser la commande `lsmeans` **avec les réserves précisées au chapitre 1, section 4.3**) ;

– les 4 commandes qui suivent construisent sur une nouvelle fenêtre graphique (commande `graphsheet()`) et après avoir découper cette fenêtre en deux (sur une même ligne), les graphiques représentant les interactions entre les deux facteurs.

Voici un extrait des résultats :

```
             Df Sum of Sq  Mean Sq  F Value     Pr(F)
      germe   1  16.70226 16.70226 41.89826 0.0000000
      trait   1   0.10378  0.10378  0.26034 0.6117603
germe:trait   1  10.01980 10.01980 25.13506 0.0000050
  Residuals 60  23.91831  0.39864

Tables of means
Grand mean           germe            trait
                 1       2        1       2
 0.63829    1.1491 0.12743  0.67856 0.59802

germe:trait Dim 1 : germe  Dim 2 : trait
                           1       2
                  1  1.5851  0.7132
                  2 -0.2280  0.4828


standard errors for differences of means
           germe    trait germe:trait
         0.15784  0.15784     0.22323
replic. 32.00000 32.00000    16.00000
```

Logiciel SAS :

En supposant les données présentes dans la table `sasuser.dents`, voici l'analyse de la variance en SAS :

```
proc glm data=sasuser.dents;
class trait germe;
```

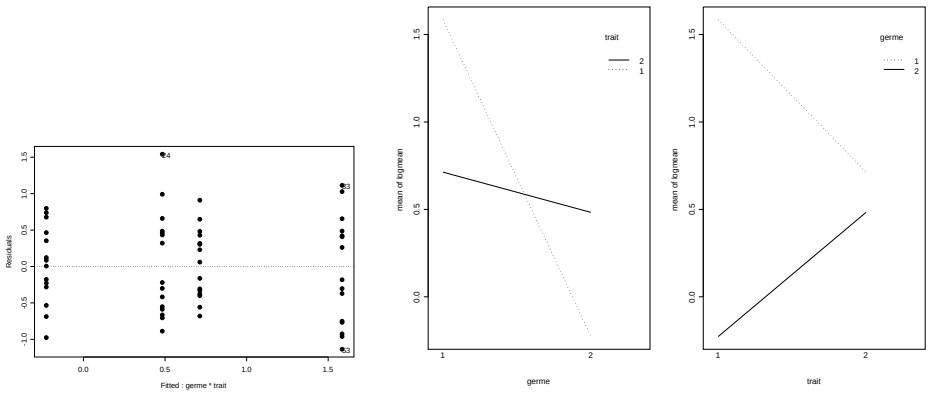


FIGURE 5.1 – Graphes 1/ à gauche, des résidus en fonction des valeurs prédites (estimateurs des différents paramètres) 2/ à droite, des interactions entre les deux facteurs (Splus)

```
model lnbac=trait germe trait*germe;
output out=sortie predicted=p student=r;
lsmeans trait germe trait*germe/out=graph;
run; quit;
proc gplot data=sortie;
plot r*p;run; quit;
proc gplot data=graph;
plot lsmean*germe=trait; run; quit;
```

Commentaires sur ces commandes :

- La seconde ligne déclare `trait` et `germe` en qualitatif.
- La troisième ligne déclare le modèle standard d'analyse de la variance à deux facteurs avec interaction. La quatrième ligne écrit, sur une table SAS, les résidus et les valeurs prédites en vue de préparer un graphique haute résolution.
- La cinquième ligne demande les moyennes ajustées (commande `lsmeans`) et les écrit dans la table "graph". Supposons que les données ne soient pas équirépétées. Le tableau d'analyse de la variance n'est plus unique, il est alors conseillé d'utiliser le tableau de type III. Le recours à la directive `lsmeans` (et non pas `means`) est nécessaire.

– Les 6 dernières commandes du programme ont pour but de tracer trois graphes qui permettent une analyse critique des résultats numériques qui suivent.

Notons que dans le cas d'un seul facteur le test de Levene est disponible de manière standard par `means .../hovtest=levene;`. Dans les cas plus compliqués il suffit de sauver les résidus, d'en calculer la valeur absolue ou le carré et de faire un nouvelle analyse de la variance sur cette dernière variable.

```
Dependent Variable: LNBAC
                                      Sum of          Mean
Source                     DF        Squares        Square   F Value     Pr > F
Model                       3     26.8258464     8.9419488     22.43     0.0001
Error                      60     23.9183125     0.3986385
Corrected Total            63     50.7441589

                   R-Square           C.V.      Root MSE           LNBAC Mean
                   0.528649       98.91739       0.63138             0.63829


Source                     DF     Type III SS     Mean Square    F Value    Pr > F
TRAIT                       1      0.10378062      0.10378062       0.26    0.6118
GERME                       1     16.70226119     16.70226119      41.90    <.0001
TRAIT*GERME                 1     10.01980464     10.01980464      25.14    <.0001

Least Squares Means

TRAIT       LNBAC        GERME        LNBAC        TRAIT   GERME         LNBAC
           LSMEAN                    LSMEAN                             LSMEAN
1      0.67855719        1       1.14914344        1       1        1.58508813
2      0.59801969        2       0.12743344        1       2       -0.22797375
                                                   2       1        0.71319875
                                                   2       2        0.48284063
```

Logiciel R :

Cette même analyse de la variance peut se faire par la suite de commandes suivantes :

```
dents=read.table("C:/Donnees/dents.txt",header=TRUE)
```

```
attach(dents)
germe=as.factor(germe)
trait=as.factor(trait)
library(car)
lm.dents=lm(LNBAC~germe:trait-1,contrasts=list(germe=contr.sum,trait=contr.sum))
summary(lm.dents)
anova(lm.dents)
Anova(lm.dents,type="II")
Anova(lm.dents,type="III")
plot(lm.dents$fit,dents.lm$res)
interaction.plot(trait,germe,LNBAC,fixed=TRUE,col=2:3,leg.bty="o")
interaction.plot(germe,trait,LNBAC,fixed=TRUE,col=2:3,leg.bty="o")
plotMeans(LNBAC,germe,trait)
library(Rcmdr)
levene.test(LNBAC,trait:germe)
```

Commentaires sur ces commandes :

- On notera tout d'abord que le logiciel R, contrairement à SAS, fait une différence entre les noms d'objets en majuscules ou en minuscules.

- Ensuite, avant toute analyse de la variance, on va s'assurer que les facteurs ont bien été définis comme des facteurs : c'est ce à quoi sont consacrées les commandes 3 et 4.

- Une fois le modèle posé, on effectue les analyses de la variance de types I, II et III (commandes 8, 9 et 10) après avoir fait appel au module `car` (commande 5). **Attention à bien préciser l'option `contrasts` dans la commande `lm` comme cela est fait, ou l'analyse de la variance de type III sera fausse.** Nous avons renoncé dans ce cas d'analyse de la variance à 2 facteurs d'obtenir directement les coefficients du modèle comme cela est fait en Splus ou en SAS ; on se contentera donc de combinaisons linéaires de ces coefficients (voir également l'exemple informatique sur les forêts du chapitre 1).

- Les instructions pour la construction des graphes constituent les 4 dernières commandes (l'instruction `plotMeans` fait la même chose que la commande `interaction.plot` avec une présentation différente).

- La dernière commande effectue enfin un test de Levene d'homoscédasticité (qui teste l'hypothèse : "Dans le modèle avec interactions la variance du bruit peut être considérée comme constante").

Voici un extrait des résultats numériques obtenus :

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
germe1:trait1   1.5851     0.1578  10.042 1.82e-14 ***
germe2:trait1  -0.2280     0.1578  -1.444  0.15386
germe1:trait2   0.7132     0.1578   4.518 2.98e-05 ***
germe2:trait2   0.4828     0.1578   3.059  0.00332 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6314 on 60 degrees of freedom
Multiple R-Squared: 0.6886,     Adjusted R-squared: 0.6679
F-statistic: 33.18 on 4 and 60 DF,  p-value: 1.36e-14

>Anova Table (Type III tests)

Response: LNBAC
             Sum Sq Df F value    Pr(>F)
(Intercept) 26.0744  1 65.4086 3.475e-11 ***
germe       16.7023  1 41.8983 1.968e-08 ***
trait        0.1038  1  0.2603    0.6118
germe:trait 10.0198  1 25.1351 5.033e-06 ***
Residuals   23.9183 60

>Levene's Test for Homogeneity of Variance

      Df F value  Pr(>F)
group  3  2.3318 0.08315 .
      60
```

Commentaires généraux : Ce jeu de données est parfaitement équilibré : pour un couple traitement-germe, il y a exactement 16 observations. Pour cette raison le tableau d'analyse de la variance de type III est identique à celui de type I (on notera que le logiciel R teste également la nullité du coefficient *intercept*). Notons également que les différents coefficients estimés des facteurs croisées peuvent être parfois accompagnés de tests de Student quant à leur potentielle nullité (comme c'est la cas dans le premier tableau en R) : **de tels tests n'ont aucune pertinence**. On voit dans ces tableaux que l'interaction est significative et que par conséquent les deux facteurs

doivent être conservés. En revanche, l'effet `trait` pris seul est non significatif, ce qui est intriguant. On comprend mieux ce qui se passe en regardant les moyennes ajustées ou les graphiques correspondants : un traitement est efficace sur un germe et le second sur l'autre. Enfin, le test de Levene montre que l'on peut, dans le cadre du modèle avec facteurs croisés, considérer la variance des erreurs comme constante.

Remarque : Dans le cas où les facteurs ont plus de deux niveaux, et dans le cas où les effets sont significatifs, une comparaison de moyennes est nécessaire. Elle peut se faire, dans le cas équilibré, suivant la méthodologie de Tukey, et dans le cas déséquilibré, le mieux est le plus souvent d'appliquer une méthode de Bonferroni à la main après avoir demandé les comparaisons deux à deux.

4.2 Analyse de la variance à deux facteurs non équirépétée

Nous prenons un exemple d'étude du temps de germination (variable `jg`), mesuré en jours, de différentes variétés (variable `var`, 3 modalités) de carottes, en fonction du type de sol (variable `sol`, 2 modalités) où ont été plantées les carottes. Ce jeu de données est extrait du livre de Searle [54]. Le nombre d'observations par type de sol et par variété y varie de 1 à 3. Les commentaires sur les différents résultats seront donnés après l'étude des commandes et résultats des trois logiciels.

Logiciel R :

Cette analyse de la variance peut se faire par la suite de commandes suivantes (**commandes qui ne demandent pas les paramètres estimés**) :

```
attach(carotte)
sol=as.factor(sol)
var=as.factor(var)
library(car)
carotte.lm=lm(jg~var*sol,contrasts=list(var=contr.sum,sol=contr.sum))
anova(carotte.lm)
Anova(carotte.lm,type="II")
Anova(carotte.lm,type="III")
```

D'où les résultats :

```
> anova(carotte.lm)
Analysis of Variance Table
```

```
Response: jg
          Df  Sum Sq Mean Sq F value   Pr(>F)
var        2  93.333  46.667  3.5000 0.075085 .
sol        1  83.901  83.901  6.2926 0.033393 *
var:sol    2 222.766 111.383  8.3537 0.008888 **
Residuals  9 120.000  13.333

> Anova(carotte.lm,type="II")
Anova Table (Type II tests)

Response: jg
           Sum Sq Df F value   Pr(>F)
var       124.734  2  4.6775 0.040475 *
sol        83.901  1  6.2926 0.033393 *
var:sol   222.766  2  8.3537 0.008888 **
Residuals 120.000  9

> Anova(carotte.lm,type="III")
Anova Table (Type III tests)

Response: jg
             Sum Sq Df  F value    Pr(>F)
(Intercept) 3497.5  1 262.3114 5.784e-08 ***
var          192.1  2   7.2048  0.013546 *
sol          123.8  1   9.2829  0.013865 *
var:sol      222.8  2   8.3537  0.008888 **
Residuals    120.0  9
```

Logiciel SAS :

L'analyse peut être faite par les commandes :

```
proc glm data=carotte;
class var sol;
model jg=var sol var*sol;
lsmeans var*sol;
run; quit;
```

Ceci donne le résultat suivant :

```
Dependent Variable: JG
                                 Sum of          Mean
Source                    DF       Squares        Square  F Value    Pr > F
Model                      5    400.000000     80.000000     6.00    0.0103
Error                      9    120.000000     13.333333
Corrected Total           14    520.000000

                  R-Square           C.V.      Root MSE            JG Mean
                  0.769231       24.34322       3.65148            15.0000

 Source                   DF       Type I SS     Mean Square  F Value    Pr > F
VAR                        2      93.3333333      46.6666667     3.50    0.0751
SOL                        1      83.9007092      83.9007092     6.29    0.0334
VAR*SOL                    2     222.7659574     111.3829787     8.35    0.0089

Source                    DF     Type III SS   Mean Square  F Value    Pr > F
VAR                        2      192.127660      96.063830     7.20    0.0135
SOL                        1      123.771429     123.771429     9.28    0.0139
VAR*SOL                    2      222.765957     111.382979     8.35    0.0089

                    Least Squares Means

                  VAR   SOL            JG
                                    LSMEAN
                  1     1       9.0000000
                  1     2      16.0000000
                  2     1      14.0000000
                  2     2      31.0000000
                  3     1      18.0000000
                  3     2      13.0000000
```

Logiciel Splus :

Voici l'exemple des carottes traité par Splus (mêmes résultats que ceux obtenus avec R ou SAS), **sans demander d'afficher les coefficients estimés** :

```
attach(carotte)
sol<-as.factor(sol)
var<-as.factor(var)
anova.lm(aov(jg~var*sol),ssType=1)
anova.lm(aov(jg~var*sol),ssType=3)
```

Commentaires généraux : Le caractère non équilibré se vérifie en comparant les sommes de carrés de type I (ou II) et III. Suivant ce qui a été décrit dans ce chapitre, **on ne prendra en considération que les sommes de carrés de type III.** Tout étant significatif, on peut consulter (en SAS) les estimations des coefficients du facteur croisé `var*sol` par la commande `lsmeans` (dans un modèle initial sans les autres coefficients estimés). Le graphique des résidus est omis, car il ne présente que peu d'intérêt compte tenu du faible effectif. L'expérience montre ainsi que le temps de germination dépend de manière complexe à la fois du type de sol et de la variété. Pour prédire ce temps de germination il faut prendre en compte chaque combinaison possible et considérer les estimations au niveau du facteur croisé telles qu'elles sont par exemple données par la directive `lsmeans` de SAS.

4.3 Analyse de la variance avec facteurs hiérarchisés

Nous présentons maintenant un jeu de données qui sera également repris dans le chapitre 10 (l'origine de ces données est le livre de Milliken et Johnson [47]). On désire comparer 8 insecticides (variable `produit`) conçus par 4 firmes différentes (variable `firme`). On suppose pour l'instant (on reviendra sur cet exemple au chapitre 10) que chaque firme produit exactement deux produits numérotés 1 et 2 pour chaque firme. Le facteur `produit` est donc hiérarchisé. On utilise pour l'expérience 24 boîtes de verre contenant du sol avec de l'herbe et 400 moustiques chacune. Pour chaque produit, on choisit sans remise 3 boîtes au hasard, on y introduit le produit, et 4 heures après, on compte le nombre de moustiques encore en vie.

Logiciel SAS :

L'analyse peut être faite par les commandes :

```
proc glm data=sasuser.insect;
class firme produit;
model nb=firme produit(firme);
means firme/tukey;
run;quit;
```

qui donne le résultat :

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	20605.33333	2943.61905	49.33	<.0001

Error	16	954.66667	59.66667
Corrected Total	23	21560.00000	

R-Square	Coeff Var	Root MSE	nb Mean
0.955720	7.022200	7.724420	110.0000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
firme	3	19971.66667	6657.22222	111.57	<.0001
produit(firme)	4	633.66667	158.41667	2.66	0.0714

Source	DF	Type III SS	Mean Square	F Value	Pr > F
firme	3	19971.66667	6657.22222	111.57	<.0001
produit(firme)	4	633.66667	158.41667	2.66	0.0714

Tukey's Studentized Range (HSD) Test for nb

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	16
Error Mean Square	59.66667
Critical Value of Studentized Range	4.04609
Minimum Significant Difference	12.759

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	firme
A	141.167	6	b
A			
A	134.667	6	a
B	91.833	6	c
C	72.333	6	d

Logiciel Splus :

Voici l'exemple des moustiques traité par Splus :

```
attach(insect)
firme<-as.factor(firme)
```

```
produit<-as.factor(produit)
anova.lm(aov(nb~firme/produit),ssType=1)
anova.lm(aov(nb~firme/produit),ssType=3)
nb.tukey<-multicomp(lm(nb~firme))
nb.tukey
```

Voici un extrait des résultats :

```
Terms added sequentially (first to last)
                   Df Sum of Sq  Mean Sq  F Value      Pr(F)
             firme  3  19971.67 6657.222 111.5736 0.00000000
produit %in% firme  4    633.67  158.417   2.6550 0.07139393
         Residuals 16    954.67   59.667

         Type III Sum of Squares
                   Df Sum of Sq  Mean Sq  F Value      Pr(F)
             firme  3  19971.67 6657.222 111.5736 0.00000000
produit %in% firme  4    633.67  158.417   2.6550 0.07139393
         Residuals 16    954.67   59.667

95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 2.7987
response variable: nb

intervals excluding 0 are flagged by '****'

    Estimate Std.Error Lower Bound Upper Bound
a-b     -6.5      5.15       -20.9         7.9
a-c     42.8      5.15        28.4        57.2 ****
a-d     62.3      5.15        47.9        76.7 ****
b-c     49.3      5.15        34.9        63.7 ****
b-d     68.8      5.15        54.4        83.2 ****
c-d     19.5      5.15         5.1        33.9 ****
```

Logiciel R :

Cette analyse de la variance peut se faire en R par la suite de commandes suivantes :

```
attach(insect)
firme=as.factor(firme)
produit=as.factor(produit)
insect.lm=lm(nb~produit:firme+firme,contrasts=list(produit=contr.sum,firme=contr.sum))
anova(insect.lm)
Anova(insect.lm,type="III")
insect.aov=aov(nb~firme)
TukeyHSD(insect.aov,ordered=TRUE)
```

D'où les résultats :

```
Response: nb
              Df  Sum Sq Mean Sq F value    Pr(>F)
firme          3 19971.7  6657.2 111.574 6.135e-11 ***
produit:firme  4   633.7   158.4   2.655    0.0714 .
Residuals     16   954.7    59.7

Anova Table (Type III tests)

Response: nb
              Sum Sq Df  F value    Pr(>F)
(Intercept)   290400  1 4867.039 < 2.2e-16 ***
firme          19972  3  111.574 6.135e-11 ***
produit:firme    634  4    2.655    0.0714 .
Residuals        955 16

  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = nb ~ firme)

$firme
        diff       lwr      upr
c-d 19.50000  5.099148 33.90085
a-d 62.33333 47.932481 76.73419
b-d 68.83333 54.432481 83.23419
a-c 42.83333 28.432481 57.23419
b-c 49.33333 34.932481 63.73419
b-a  6.50000 -7.900852 20.90085
```

Commentaires : En comparant les sommes de carrés de type I et III, nous vérifions que le dispositif est équilibré. L'effet hiérarchisé est plutôt non significatif, mais on doit garder à l'esprit que l'on n'est pas loin de la limite des 5%. L'effet de la firme est par contre clairement significatif, les aptitudes des firmes à produire des insecticides performants sont clairement différentes. Par soucis de concision, on a omis les graphes de résidus ainsi que les moyennes.
De plus, comme le dispositif est équilibré et que l'utilisateur le sait, on peut utiliser des comparaisons des moyennes obtenues par la méthode de Tukey. Les résultats affichés par SAS ne sont pas présentés de la même manière que ceux affichés par Splus ou R, mais leur signification est la même : les moyennes des firmes `a` et `b` ne sont pas significativement différentes. En conclusion l'expérience ne montre pas de différences entre les produits des différentes firmes. Elle montre en revanche une différence entre les firmes `c`, `d` et le groupe (`a` et `b`).

5 Exercices

Exercice 5.1

(*) Soit le jeu de données suivant pour deux facteurs à deux niveaux (données totalement inventées pour que les calculs tombent juste) :

facteur 1	1	1	1	1	1	2	2	2
facteur 2	1	1	1	2	2	1	1	2
réponse	19	15	14	10	6	9	11	6

Calculez les estimateurs dans le cas des paramétrisations (5.1) et (5.2). On calculera plus rapidement si on représente les données dans un tableau 2×2.

Exercice 5.2

(*) [Hétéroscédasticité] Soit un modèle d'analyse de la variance à un facteur, vérifiant les postulats **P1, P3** et **P4** et tel que pour les I modalités du facteur on ait $\text{Var}(\varepsilon_{ij}) = \sigma_i^2$ pour tout $j = 1, \ldots, n_i$. La variance des erreurs dépend donc de la modalité considérée. Déterminer alors les estimateurs $\widehat{\mu}_i$ et $\widehat{\sigma}_i^2$ par maximum de vraisemblance des différents paramètres du modèle (soit $2I$ paramètres). Si on suppose maintenant que $\mu_i = \mu$ pour $i = 1, \ldots, I$, que valent alors les $\widehat{\sigma}_i^2$ et $\widehat{\mu}$?

Exercice 5.3

(**) Montrer que $H_0^{(1)}$ est équivalente à $\theta_{i.} = (cte)$ dans le modèle (5.1). Montrer que $H_0^{(3)}$ est équivalente à $\forall (i,j) \neq (i',j') \in \{1, \ldots, I\} \times \{1, \ldots, J\} : \theta_{ij} - \theta_{i'j} - \theta_{ij'} + \theta_{i'j'} = 0$.

Exercice 5.4

(**) Pour introduire des effets différentiels dans un modèle d'analyse de la variance à **un** facteur,

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik},$$

quel type de contrainte doit-on utiliser ?

$$\sum_{i=1}^{I} \alpha_i = 0 \text{ ou bien } \sum_{i=1}^{I} n_i \alpha_i = 0 \text{ ?}$$

Eléments de solution : Par cohérence avec la décomposition marginale utilisée dans le cas où il y a deux facteurs, on serait tenté de poser la première contrainte. Ce n'est pas la bonne réponse. En effet :

- *L'estimation et donc la définition précise du paramètre μ importe peu, puisque ce paramètre de moyenne générale n'est pas utilisé. N'oublions pas que le but d'une expérience est de comparer ; toute l'attention est donc focalisée sur les effets différentiels α_i. Le premier type de contraintes a peu d'intérêt.*

- *Le second type n'a d'autre intérêt que calculatoire. Si on l'utilise, l'estimateur $\widehat{\mu}$ de μ est la moyenne générale : $Y_{..}$ (estimateur sous l'hypothèse nulle), nécessaire pour la construction du test de Fisher. En particulier, la formule $SC = \sum_{ik}(Y_{i.} - Y_{..})^2 = \sum_{ik}(\widehat{\alpha}_i)^2$ ne serait pas vraie sinon.*

Cette réponse est en contradiction avec celle que l'on fait pour deux facteurs. Nous voyons donc une fois de plus sur cet exemple, l'intérêt de travailler avec des modèles réguliers, comme l'est le modèle de notre première présentation.

Exercice 5.5

(**) [Décomposition de type I] Soit un modèle linéaire et supposons que l'on ait scindé le paramètre θ en différents sous-ensembles $\theta_1, \cdots, \theta_m$. Une telle décomposition est appelée une *partition*. Une partition naturelle est la décomposition en effet principaux et interaction dans le modèle à deux facteurs croisés (5.2). En toute rigueur, ce modèle est non-régulier et le lecteur pourra consulter le chapitre 7. La partition fait que l'on peut écrire

$$Y = X_1 \cdot \theta_1 + \cdots + X_m \cdot \theta_m + \varepsilon.$$

Notez bien que l'ordre importe dans l'écriture du mode et que nous supposerons toujours que les effets principaux sont avant les interactions, que les interactions doubles sont avant les triples, etc... De manière générale, on définit les espaces V_i $i = 1, \cdots, m$ de la façon suivante :

$$V_i \text{ est l'orthogonal de } [X_1, \cdots, X_{i-1}] \text{ dans } [X_1, \cdots, X_i].$$

On définit le test associé au i-ème élément du modèle comme le test de l'hypothèse nulle :

$$P_{[V_i]}X \cdot \theta = 0.$$

On considère désormais un modèle d'analyse de la variance à deux facteurs et on désire pondérer la décomposition par les effectifs :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

avec maintenant les contraintes suivantes : $\forall j,\ \sum_i n_{ij}\gamma_{ij} = 0,\ \ \forall i,\ \sum_j n_{ij}\gamma_{ij} = 0$, et également $\sum_i n_{i+}\alpha_i = 0,\ \ \sum_j n_{+j}\beta_j = 0$ (le + veut dire la somme sur l'indice qu'il remplace). Montrer que l'hypothèse $\forall i,\ \alpha_i = 0$ s'écrit dans le modèle (5.1) sous la forme : "$h_i := \sum_j n_{ij}\theta_{ij}$, ne dépend pas de i". Montrer que le test de cette hypothèse est obtenu par la décomposition de type I.

Exercice 5.6

(**) Nous allons illustrer une nouvelle fois l'abominable complexité de l'option `solution` de SAS en analyse de la variance à deux facteurs. Voici un exemple volontairement simple et dont les données ont été inventées. L'utilisation de `solution` donne la valeur -12 à la fin du tableau pour $a * b = (1, 1)$. Comment s'interprète t-elle ?

	a	1	1	1	1	1	2	2	2
Les données sont	b	1	1	2	2	2	1	2	2
	Y	5	3	25	27	32	12	12	21

```
proc glm ;
class a b;
model y= a b a*b/solution;
lsmeans a b a*b;run; quit;
```

Voici un extrait des résultats numériques obtenus :

```
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        3     792.0000000     264.0000000      22.96    0.0055
Error                        4      46.0000000      11.5000000
Corrected Total              7     838.0000000

           R-Square     Coeff Var      Root MSE        y Mean

           0.945107      17.84824      3.391165      19.00000

Source                      DF     Type III SS     Mean Square    F Value    Pr > F
```

a	1	6.8571429	6.8571429	0.60	0.4831
b	1	555.4285714	555.4285714	48.30	0.0023
a*b	1	61.7142857	61.7142857	5.37	0.0814

Parameter			Estimate		Standard Error	t Value	Pr > |t|
Intercept			24.00000000	B	2.39791576	10.01	0.0006
a	1		4.00000000	B	3.09569594	1.29	0.2659
a	2		0.00000000	B	.	.	.
b	1		-12.00000000	B	4.15331193	-2.89	0.0446
b	2		0.00000000	B	.	.	.
a*b	1	1	-12.00000000	B	5.18009009	-2.32	0.0814
a*b	1	2	0.00000000	B	.	.	.
a*b	2	1	0.00000000	B	.	.	.
a*b	2	2	0.00000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

a	y LSMEAN
1	16.0000
2	18.0000

b	y LSMEAN
1	8.0000
2	26.0000

a	b	y LSMEAN
1	1	4.0000
1	2	28.0000
2	1	12.0000
2	2	24.0000

Chapitre 6

Analyse de la covariance

Dans ce chapitre, nous montrons comment mélanger, dans un modèle linéaire, des variables qualitatives et quantitatives. Nous introduisons en particulier le modèle avec hétérogénéité des pentes qui définit, dans une certaine mesure, une interaction entre variable quantitative et qualitative.

1 Le modèle

Dans de nombreuses situations, l'ensemble des variables explicatives se compose à la fois de variables quantitatives et qualitatives. Nous n'avons pour l'instant présenté que des techniques travaillant dans un cas (quantitatif : régression linéaire) ou dans l'autre (qualitatif : analyse de la variance). Comme ces deux techniques découlent toutes deux du modèle linéaire, il est possible en quelque sorte de mélanger les genres : c'est ce que l'on appelle l'analyse de la covariance (ce nom a une raison historique, mais n'est pas vraiment en adéquation avec la méthode elle-même).

1.1 Un exemple

Commençons par l'exemple suivant : supposons que l'on veuille expliquer la taille de fillettes de 6 à 10 ans en fonction de leur âge. Pour une fillette donnée, dans la plage d'âges considérée, un modèle de régression linéaire est raisonnable. En revanche, il est bien connu qu'il y a des individus plus grands que d'autres et des individus dont la taille va augmenter plus vite. Si on posait un modèle de régression unique, les postulats du modèle linéaire ne seraient pas respectés : par exemple tous les termes ε_i associés à un individu grand auraient tendance à être positifs (et le premier postulat $E(\varepsilon_i) = 0$ ne serait plus vérifié). Il faut donc que les paramètres de la régression changent d'un

individu à un autre. C'est donc le modèle avec "hétérogénéité des pentes" (qui est le modèle le plus complexe) que l'on va poser.

Soit Y_{ij} la taille de l'individu i à l'âge numéro j et soit age_j cet âge, (age_1 pourra être par exemple 6 ans), avec $i = 1, \cdots, I$ et $j = 1, \cdots, J$. On suppose que Y_{ij} vérifie le modèle linéaire suivant :

$$Y_{ij} = \mu_i + \beta_i \cdot age_j + \varepsilon_{ij}.$$

Deux questions statistiques se posent :

i. Est-ce que les β_i sont différents ? S'ils le sont, on dira alors que l'on a hétérogénéité des pentes ;

ii. Est-ce que les μ_i sont différents ? S'ils le sont, on dira alors que l'on a hétérogénéité des constantes.

On peut introduire des effets différentiels, soit pour tout $i = 1, \cdots, I$:

$$\mu_i = \mu + \alpha_i \quad \text{et} \quad \beta_i = \beta + \gamma_i$$

avec les conditions habituelles

$$\sum_i \gamma_i = \sum_i \alpha_i = 0.$$

On obtient alors le modèle tel que pour $i = 1, \cdots, I$ et $j = 1, \cdots, J$,

$$Y_{ij} = \mu + \alpha_i + \beta \cdot age_j + \gamma_i \cdot age_j + \varepsilon_{ij}, \tag{6.1}$$

modèle qui est strictement équivalent au premier. Le dernier terme $\gamma_i \cdot age_j$ qui dépend des deux variables peut être considéré comme une interaction entre le facteur `individu` et la variable quantitative `age`. La notation des logiciels informatiques du modèle 6.1 est d'ailleurs :

```
taille = individu + age + age * individu.
```

On répond d'abord à la question i. en testant l'hypothèse :

- pour tout $i = 1, \cdots, I,\ \gamma_i = 0.$

Si cet effet est non significatif (donc si le test ne rejette pas l'hypothèse $\gamma_i = 0$ pour tout $i = 1, \cdots, I$), on peut alors répondre à la question ii. en testant l'hypothèse :

- pour tout $i = 1, \cdots, I,\ \alpha_i = 0$.

Dans ce dernier cas, le facteur `individu` disparaît complètement du modèle. Remarquons que si l'on admet seulement la nullité des γ_i, le modèle (6.1) devient

$$Y_{ij} = \mu + \alpha_i + \beta \cdot \text{age}_j + \varepsilon_{ij}$$

qui revient à rajouter la simple covariable âge au modèle d'analyse de la variance à un facteur : `individu`.

1.2 Le modèle général

Dans les cas les plus généraux, on peut avoir plusieurs facteurs avec une structure croisée ou hiérarchique ainsi que plusieurs variables intervenant de manière linéaire, polynômiale ou de façon plus complexe encore. Voyons maintenant la difficulté d'écrire un tel modèle sous une forme générale.

Supposons que l'on ait une variable à expliquer Y, observée pour n individus (soit Y_i, $i = 1, \cdots, n$), variable dépendant a priori de p variables quantitatives (soit $V^{(1)}, \cdots, V^{(p)}$) et de q facteurs qualitatifs (soit $F^{(1)}, \cdots, F^{(q)}$), le facteur $F^{(j)}$ pouvant prendre m_j modalités $mod_k^{(j)}$). On peut remarquer que faire l'analyse de la variance sur les facteurs $F^{(j)}$ revient au même, en fait, que de faire la régression sur les variables indicatrices des classes $F^{(j,k)} = \mathbb{1}_{\{F^{(j)} = mod_k^{(j)}\}}$. Supposons maintenant que l'on s'en tienne à des termes d'interaction d'ordre 2 entre ces différentes variables et facteurs et que tous les facteurs soient croisés. Le modèle s'écrit sous la forme "abrégée" des logiciels :

$$Y = \sum_{1 \le j \le p} V^{(j)} + \sum_{1 \le j \le q} F^{(j)} + \sum_{1 \le j \le k \le p} V^{(j)} V^{(k)} + \sum_{1 \le j \le k \le q} F^{(j)} F^{(k)} + \sum_{1 \le j \le p,\ 1 \le k \le q} V^{(j)} F^{(k)} + \varepsilon_i.$$

Ouf ! Et pourtant nous nous en sommes seulement tenus aux termes croisés d'ordre 2. Si l'on retranscrit ce qui précède en termes mathématiques, cela s'écrira :

$$\begin{aligned} Y_i = &\sum_{1 \le j \le p} \theta_j \cdot V_i^{(j)} + \sum_{1 \le j \le q} \sum_{1 \le j' \le m_j} \mu_{jj'} \cdot F_i^{(j,j')} + \sum_{1 \le j \le k \le p} \theta_{jk} \cdot V_i^{(j)} V_i^{(k)} \\ &+ \sum_{1 \le j \le k \le q} \sum_{1 \le j' \le m_j} \sum_{1 \le k' \le m_k} \mu_{jj'kk'} \cdot F_i^{(j,j')} F_i^{(k,k')} + \sum_{1 \le j \le p} \sum_{1 \le k \le q} \sum_{1 \le k' \le m_k} \nu_{jkk'} \cdot V_i^{(j)} F_i^{(k,k')} + \varepsilon_i, \end{aligned} \tag{6.2}$$

pour $i = 1, \cdots, n$, avec (ε_i) les erreurs du modèle et $\theta_j, \mu_{jj'}, \theta_{jk}, \mu_{jj'kk'}, \nu_{jkk'}$ des constantes. On s'aperçoit ici concrètement de la difficulté d'écrire explicitement le

modèle. Cependant, il est toujours possible d'écrire un tel modèle sous la forme simple du modèle linéaire (3.5) du chapitre 3. On aboutit alors à un modèle linéaire liant Y et les différentes variables quantitatives $Z^{(i)}$: c'est bien le cadre du modèle (3.5).

Un modèle d'analyse de la covariance s'écrit donc comme un modèle linéaire classique et ne présente pas plus de difficultés techniques que celles évoquées au cours des chapitres précédents. Son interprétation devra plutôt suivre la démarche suivante : faire des régressions intra-groupes et faire si possible apparaître des effets différentiels inter-groupes. Voyons ceci sur l'exemple initial...

2 Exemple traité par logiciels informatiques

On a mesuré (en cm) et année par année entre 6 à 10 ans (variable `age`) la taille (variable `taille`) de 7 fillettes (chaque variable étant une modalité du facteur `ind`) :

ind	6 ans	7 ans	8 ans	9 ans	10 ans
1	116	122	126.6	132.6	137.6
2	117.6	123.2	129.3	134.5	138.9
3	121	127.3	134.5	139.9	145.4
4	114.5	119	124	130	135.1
5	117.4	123.2	129.5	134.5	140
6	113.7	119.7	125.3	130.1	135.9
7	113.6	119.1	124.8	130.8	136.3

(les données sont extraite du livre de Tanner [57]). La période considérée, après la petite enfance et avant la poussée pubertaire, est une période pendant laquelle la croissance en taille d'êtres humains est presque linéaire. Le modèle d'hétérogénéité des pentes est donc bien adapté. Pour pouvoir interpréter les constantes (ou `intercept`) comme des tailles à l'âge moyen et non pas comme des tailles à l'âge zéro, on a centré la variable `age` en définissant `agec` qui est l'âge centré c'est-à-dire la différence d'âge par rapport à 8 ans.

Logiciel SAS :

Pour analyser ces données, on lance trois appels de `proc glm` :

```
proc glm data=...;
```

```
class ind;
model taille=agec ind agec*ind; run; quit;
proc glm data=...;
class ind;
model taille=age ind age*ind; run; quit;
proc glm data=...;
class ind;
model taille=agec*ind/solution noint; run; quit;
```

La troisième commande a pour but de définir un modèle régulier, ce qui permet d'obtenir des estimateurs interprétables sous forme de taille moyenne et de vitesse de croissance. Les tests obtenus par les deux premières commandes feront l'objet d'une interprétation commune aux 3 logiciels. Voici les résultats de ces commandes :

```
                                 Sum of
Source                    DF        Squares     Mean Square    F Value    Pr > F
Model                     13     2491.474429      191.651879     981.39    <.0001
Error                     21        4.101000        0.195286
Corrected Total           34     2495.575429

                  R-Square     Coeff Var      Root MSE    taille Mean
                  0.998357      0.346566      0.441911       127.5114

-------------------------------------------------------------------------------
Source                    DF       Type I SS     Mean Square    F Value    Pr > F
age                        1     2169.515571     2169.515571    11109.4    <.0001
ind                        6      316.459429       52.743238     270.08    <.0001 (A)
age*ind                    6        5.499429        0.916571       4.69    0.0035

Source                    DF     Type III SS     Mean Square    F Value    Pr > F
age                        1     2169.515571     2169.515571    11109.4    <.0001
ind                        6        4.253299        0.708883       3.63    0.0125 (B)
age*ind                    6        5.499429        0.916571       4.69    0.0035

-------------------------------------------------------------------------------
Source                    DF       Type I SS     Mean Square    F Value    Pr > F
agec                       1     2169.515571     2169.515571    11109.4    <.0001
ind                        6      316.459429       52.743238     270.08    <.0001 (C)
agec*ind                   6        5.499429        0.916571       4.69    0.0035

Source                    DF     Type III SS     Mean Square    F Value    Pr > F
```

```
agec                  1     2169.515571     2169.515571    11109.4    <.0001
ind                   6      316.459429       52.743238      270.08    <.0001 (C)
agec*ind              6        5.499429        0.916571        4.69    0.0035

-------------------------------------------------------------------------------

                                          Standard
          Parameter           Estimate         Error    t Value    Pr > |t|
          ind       1      126.9600000    0.19762880     642.42      <.0001
          ind       2      128.7000000    0.19762880     651.22      <.0001
            :                    :             :             :           :
          agec*ind  1        5.3800000    0.13974467      38.50      <.0001
          agec*ind  2        5.3900000    0.13974467      38.57      <.0001
            :                    :             :             :           :
```

Logiciel Splus :

Cette analyse de la covariance peut se faire en Splus par la suite de commandes suivantes (**attention à prendre avec précaution les coefficients estimés**) :

```
ind<-as.factor(ind)
summary(lm(taille~age*ind))
anova.lm(aov(taille~age*ind),ssType=3)
anova.lm(aov(taille~agec*ind),ssType=3)
menuAov(taille~agec*ind,means.p=T,lsmeans.p=T)
```

D'où les résultats (extraits) :

```
Type III Sum of Squares
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
       age  1  2169.516 2169.516 11109.44 0.00000000
       ind  6     4.253    0.709     3.63 0.01252009 (B)
  age:ind  6     5.499    0.917     4.69 0.00354688
Residuals 21     4.101    0.195

Type III Sum of Squares
          Df Sum of Sq  Mean Sq  F Value       Pr(F)
      agec  1  2169.516 2169.516 11109.44 0.000000000
       ind  6   316.459   52.743   270.08 0.000000000 (C)
 agec:ind  6     5.499    0.917     4.69 0.003546876
Residuals 21     4.101    0.195
Estimated effects are balanced
```

```
Tables of means
Grand mean

 127.51

 agec
    -2     -1      0      1      2
 116.38 121.94 127.51 133.08 138.65

 ind
     1     2      3      4      5      6      7
 126.96 128.7 133.62 124.52 128.92 124.94 124.92

 agec:ind
Dim 1 : agec
Dim 2 : ind
        1      2      3      4      5      6      7
-2 116.20 117.92 121.34 114.08 117.62 113.98 113.50
-1 121.58 123.31 127.48 119.30 123.27 119.46 119.21
 0 126.96 128.70 133.62 124.52 128.92 124.94 124.92
 1 132.34 134.09 139.76 129.74 134.57 130.42 130.63
 2 137.72 139.48 145.90 134.96 140.22 135.90 136.34
```

Logiciel R :

Voici enfin la même analyse à l'aide de R :

```
ind=as.factor(ind)
library(car)
fille.new=groupedData(taille~age|ind,data=fille)
plot(fille.new)
anova(lm(taille~age*ind))
Anova(lm(taille~age*ind),type="III",contrasts=list(age=contr.sum,ind=contr.sum))
anova(lm(taille~agec*ind))
summary(lm(taille~(agec:ind-1)+ind))
```

D'où les résultats (extraits) :

```
Response: taille
```

```
          Df  Sum Sq Mean Sq    F value    Pr(>F)
age        1 2169.52 2169.52 11109.4433 < 2.2e-16 ***
ind        6  316.46   52.74   270.0824 < 2.2e-16 *** (A)
age:ind    6    5.50    0.92     4.6935  0.003547 **
Residuals 21    4.10    0.20

Anova Table (Type III tests)

Response: taille
             Sum Sq Df   F value    Pr(>F)
(Intercept) 1067.06  1 5464.0736 < 2.2e-16 ***
age          289.44  1 1482.1565 < 2.2e-16 ***
ind            4.25  6    3.6300  0.012520 *     (B)
age:ind        5.50  6    4.6935  0.003547 **
Residuals      4.10 21

Response: taille
          Df  Sum Sq Mean Sq    F value    Pr(>F)
agec       1 2169.52 2169.52 11109.4433 < 2.2e-16 ***
ind        6  316.46   52.74   270.0824 < 2.2e-16 *** (C)
agec:ind   6    5.50    0.92     4.6935  0.003547 **
Residuals 21    4.10    0.20

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
ind1      126.9600     0.1976  642.42   <2e-16 ***
ind2      128.7000     0.1976  651.22   <2e-16 ***
  :           :          :       :         :
agec:ind1   5.3800     0.1397   38.50   <2e-16 ***
agec:ind2   5.3900     0.1397   38.57   <2e-16 ***
  :           :          :       :         :
```

Commentaires : L'ajustement est excellent : non seulement le coefficient R^2 est de $\simeq 0.998$, mais en plus l'écart-type est de 4 mm, ce qui paraît inespéré compte tenu des erreurs de mesure non compressibles. Tous les effets sont significatifs : il était tout à fait évident que la taille dépendait de l'âge. L'analyse montre également que la vitesse de croissance depend de l'individu par la variable `ind*age`. En ce qui concerne les tests de l'effet `ind` trois tests sont proposés :

– le test de type I avec `age` (A) ;

- le test de type III avec `age` (B) ;
- les tests de type I ou III avec `agec` (C)(ils sont équivalents car ce modèle est orthogonal, voir Chapitre 7).

On vérifie que (A) et (C) sont identiques, ils testent finalement l'hypothèse nulle : "les tailles à l'âge moyen sont égales". C'est ce test qui est le plus interprétable. (B) teste l'hypothèse nulle : "les tailles extrapolées à l'âge 0 sont égales". Ce dernier test a peu de sens car, en début de vie, la croissance de l'être humain est non linéaire et donc la taille extrapolée n'apporte pas vraiment d'information. **On voit ici un des rares intérêts de la décomposition I qui permet d'obtenir le test pertinent sans centrer**. Avec les trois logiciels, on a demandé l'affichage des estimations des coefficients, ce qui permet de lire les tailles moyennes à 8 ans par individu, ainsi que les vitesses de croissance (par individu).

Conclusion générale : L'expérience montre que dans la plage d'âges considérée, la croissance des fillettes est très convenablement modélisée par une croissance linéaire (ce point est le seul élément non trivial). Pour le reste, la taille augmente avec l'âge, la taille moyenne dépend de l'individu ainsi que la vitesse de croissance.

Chapitre 7

Modèles non réguliers et orthogonalité

Dans ce chapitre, on donne des éléments pour étudier un modèle linéaire dit non-régulier et pour utiliser la notion d'orthogonalité qui permet de simplifier le modèle.

1 Cas de modèles non réguliers

Certains modèles ne peuvent être paramétrés de façon régulière : ils sont naturellement sur-paramétrés. Un exemple simple est celui du modèle additif en analyse de la variance à 2 facteurs. Considérons le cas où les 2 facteurs ont chacun 2 niveaux et que les 4 combinaisons sont observées une fois et une seule. On a donc, avec les notations vues précédemment :

$$\begin{aligned} Y_{11} &= \mu + a_1 + b_1 + \varepsilon_{11} \\ Y_{12} &= \mu + a_1 + b_2 + \varepsilon_{12} \\ Y_{21} &= \mu + a_2 + b_1 + \varepsilon_{21} \\ Y_{22} &= \mu + a_2 + b_2 + \varepsilon_{22}. \end{aligned}$$

La matrice X du modèle vaut :

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

On voit donc que tout vecteur de la forme $(\alpha+\beta, -\alpha, -\alpha, -\beta, -\beta)$ donne la valeur zéro lorsqu'il est multiplié par la matrice X. Les valeurs $\mu, a_i, b_i,\ (i=1,2)$ ne sont donc pas identifiables de manière unique.

Définition 7.1 *Le modèle est dit non régulier quand la matrice X est non injective c'est-à-dire s'il existe $\theta \neq 0$ tel que $X \cdot \theta = 0$.*

On note $K := Ker(X) := \{z \in \mathbb{R}^n,\ X \cdot z = 0\}$ le noyau de X. Commençons par deux remarques :

- $X \cdot \widehat{\theta}$ reste unique, puisque ce vecteur est la projection de Y sur $[X]$,
- $\widehat{\theta}$ ne peut être unique puisque si $\widehat{\theta}$ est solution et si $z \in K$, $\widehat{\theta} + z$ est encore solution. Compte tenu de (i), si $\widehat{\theta}$ est une solution particulière, l'ensemble des solutions s'écrit $\{\widehat{\theta} + z,\ z \in K\}$.

Ceci nous amène à la définition suivante :

Définition 7.2 *Soit M une matrice, M^- est une matrice pseudo-inverse de M si $M \cdot M^- \cdot M = M$.*

Proposition 7.1 *Si $(X' \cdot X)^-$ est une matrice pseudo-inverse de $X' \cdot X$, alors $\widehat{\theta} = (X' \cdot X)^- X' \cdot Y$ est une solution des équations normales*

$$(X' \cdot X) \cdot \widehat{\theta} = X' \cdot Y.$$

Démonstration : On sait que $P_{[X]}Y$ existe de manière unique, donc il existe $u \in \mathbb{R}^k$ tel que $X' \cdot Y = X' \cdot P_{[X]}Y = X' \cdot X \cdot u$. Soit $\widehat{\theta} = (X' \cdot X)^- X' \cdot Y$. Alors, $\widehat{\theta}$ vérifie les équations normales puisque :

$$\begin{aligned} X' \cdot X \cdot \widehat{\theta} &= (X' \cdot X)(X' \cdot X)^- X' \cdot Y \\ &= (X' \cdot X)(X' \cdot X)^- X' \cdot X \cdot u \\ &= X' \cdot X \cdot u = X' \cdot Y \end{aligned}$$

d'après la définition de la matrice pseudo-inverse de $X' \cdot X$. ■

Parmi toutes les matrices pseudo-inverses qui conduisent à une solution particulière des équations normales, certaines sont plus intéressantes que d'autres. C'est ce que nous allons maintenant étudier.

Contraintes d'identifiabilité

Proposition 7.2 *On suppose que* $rg(X) = dim[X] = h < k$ *de sorte qu'il y ait* $(k-h)$ *paramètres redondants. On définit une matrice* K *à* $(k-h)$ *lignes et* k *colonnes que l'on suppose de plein rang et telle que :*

$$Ker(K) \cap Ker(X) = \{0\}.$$

Alors

- *La matrice* $(X' \cdot X + K' \cdot K)$ *est inversible et son inverse est une matrice pseudo-inverse de* $(X' \cdot X)$.
- *Le vecteur* $\widehat{\theta} = (X' \cdot X + K' \cdot K)^{-1} X' \cdot Y$ *est l'unique solution des équations normales qui vérifie* $K \cdot \widehat{\theta} = 0$.

Démonstration : Les propriétés de dimension d'espaces vectoriels impliquent qu'il existe un seul $\widehat{\theta}$ tel que

$$X \cdot \widehat{\theta} = P_{[X]} \cdot Y \quad \text{avec} \quad K \cdot \widehat{\theta} = 0.$$

Considérons le problème de la minimisation en θ de $\|Y - X \cdot \theta\|^2 + \|K \cdot \theta\|^2$. La valeur $\widehat{\theta}$ précédente est bien la solution de ce problème de minimum car elle minimise séparément les deux termes positifs. Le problème précédent peut alors se mettre sous la forme

$$\left\| \left(\frac{Y}{0} \right) - \left(\frac{X}{K} \right) \cdot \theta \right\|^2 \text{ est minimum,}$$

la barre horizontale désignant la concaténation de matrices, c'est-à-dire la mise l'une sous l'autre de chacune des colonnes des matrices. La matrice de droite est de plein rang car

$$\begin{aligned} \left(\frac{X}{K} \right) \cdot \theta = 0 \quad &\Rightarrow \quad X \cdot \theta = 0 = K \cdot \theta \\ &\Rightarrow \quad \theta \in \mathrm{Ker}([K]) \cap \mathrm{Ker}([X]) \Rightarrow \theta = 0. \end{aligned}$$

Nous savons alors que la solution des moindres carrés est donnée par

$$\widehat{\theta} = \left(\left(\frac{X}{K} \right)' \cdot \left(\frac{X}{K} \right) \right)^{-1} \cdot \left(\frac{X}{K} \right)' \cdot \left(\frac{Y}{0} \right) = (X' \cdot X + K' \cdot K)^{-1} X' \cdot Y.$$

Il reste à montrer que $(X' \cdot X + K' \cdot K)^{-1}$ est une matrice pseudo-inverse de $(X' \cdot X)$, ce qui se déduit facilement, puisque :

$$(X' \cdot X)(X' \cdot X + K' \cdot K)^{-1}(X' \cdot X) = X' \cdot P_{[X]} \cdot X = X' \cdot X$$

car, par définition, $P_{[X]} \cdot X = X$. ■

Exemple 7.1 *Soit le modèle suivant d'analyse de la variance à un facteur*

$$Y_{i,j} = \mu_i + \varepsilon_{i,j} \quad pour \quad i = 1, \cdots, 4 \quad et \quad j = 1, 2.$$

Ce modèle est régulier de dimension 4*, mais si on le paramétrise avec des effets différentiels :*

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j} \quad pour \quad i = 1, \cdots, 4 \quad et \quad j = 1, 2,$$

alors ce modèle n'est plus régulier : on a $p = 5$ *et* $k = 4$*. La contrainte* $\sum_{i=1}^{4} \alpha_i = 0$ *rendra le modèle identifiable.*

Exemples de contrainte : l'opérateur de balayage ou "sweep operator"

Ce type de contraintes est très important en analyse de la variance car il permet de comprendre le "pourquoi" de certaines sorties de programmes informatiques. Lorsque l'on introduit informatiquement un modèle linéaire, le programme doit "traquer" les colinéarités entre les colonnes de la matrice X. S'il détecte une ligne colinéaire aux précédentes, elle est supprimée. Cela permet de diminuer informatiquement la taille du modèle. Mais cela revient aussi en se plaçant dans le cadre précédent à imposer des contraintes de nullité de certaines coordonnées du vecteur θ. Il s'agit d'un balayage des colonnes de la matrice X dans l'ordre d'entrée. Ceci explique pourquoi l'ordre des termes peut avoir son importance (bien que l'addition soit commutative) et pourquoi on trouve en analyse de la variance le dernier niveau d'un facteur fixé à zéro. Ainsi, dans notre exemple introductif d'analyse de la variance à deux facteurs, "le sweep operator" introduira les contraintes $a_2 = b_2 = 0$, ce qui n'est pas symétrique et nuit le plus souvent à l'interprétation.

Fonctions estimables et contrastes

Il existe des fonctions de θ qui ne dépendent pas de la solution particulière des équations normales, c'est-à-dire du type de contraintes d'identifiabilité choisi. Ces fonctions sont appelées estimables car elles sont intrinsèques.

Définition 7.3 *Une combinaison linéaire* $C' \cdot \theta$ *est dite estimable si elle ne dépend pas du choix particulier d'une solution des équations normales. On caractérise ces fonctions comme étant celles qui s'écrivent* $C' \cdot \theta = D' \cdot X \cdot \theta$*, où* D *est une matrice de plein rang.*

Définition 7.4 *En analyse de la variance, on définit les "contrastes" comme les combinaisons linéaires de poids nul :* $C' \cdot \theta$ *avec* $C' \cdot \mathbb{1} = 0$*.*

Si on reprend encore notre exemple introductif d'analyse de la variance à deux facteurs, $a_1 - a_2$ est un contraste. Les contrastes sont le plus souvent des fonctions estimables.

2 Orthogonalité pour des modèles réguliers

L'orthogonalité est une notion qui peut notablement simplifier la résolution et la compréhension d'un modèle linéaire. Un modèle linéaire admet le plus souvent une décomposition naturelle des paramètres θ (voir les deux exemples ci-dessous) et conséquemment une décomposition de la matrice X associée au modèle. On va s'intéresser ici à l'orthogonalité éventuelle des différents espaces associés à cette décomposition (l'orthogonalité sera toujours comprise par la suite au sens d'orthogonalité liée au produit scalaire euclidien usuel). Le problème sera plus ou moins délicat suivant que le modèle est régulier ou non. En premier lieu, illustrons par deux exemples ce que l'on entend par décomposition des paramètres :

Exemple 7.2 *Soit le modèle de régression linéaire multiple sur trois variables $Z^{(1)}, Z^{(2)}$ et $Z^{(3)}$:*

$$Y_i = \mu + \beta_1 Z_i^{(1)} + \beta_2 Z_i^{(2)} + \beta_3 Z_i^{(3)} + \varepsilon_i \quad , \quad i = 1, \cdots , \quad n > 4.$$

Le vecteur θ comprend 4 coordonnées : $\mu, \beta_1, \beta_2, \beta_3$ et la matrice X quatre colonnes. Assez naturellement ici, on peut considérer la décomposition, plus précisément on parlera par la suite de partition, en quatre éléments. La partition de la matrice revient alors à l'écrire comme concaténation de 4 vecteurs colonnes. L'orthogonalité de la partition correspondra alors strictement à l'orthogonalité des 4 droites vectorielles : $[\mathbb{1}]$, $[Z^{(1)}]$, $[Z^{(2)}]$ *et* $[Z^{(3)}]$.

Exemple 7.3 *Soit le modèle de régression quadratique sur deux variables $Z^{(1)}$ et $Z^{(2)}$*

$$Y_i = \mu + \beta_1 Z_i^{(1)} + \beta_2 Z_i^{(2)} + \gamma_1 \left(Z_i^{(1)}\right)^2 + \gamma_2 \left(Z_i^{(2)}\right)^2 + \delta Z_i^{(1)}.Z_i^{(2)} + \varepsilon_i \quad , \quad i = 1, \cdots , n > 6.$$

Ici, plutôt que de demander comme précédemment l'orthogonalité de chacun des régresseurs (ce qui serait beaucoup demander...), on peut définir la partition naturelle correspondant à :

- *la constante μ ;*
- *les effets linéaires β_1, β_2 ;*
- *les effets carrés γ_1, γ_2 ;*
- *les effets produits δ.*

L'orthogonalité de la partition est alors définie comme l'orthogonalité des sous-espaces vectoriels : $[\mathbb{1}]$, $[(Z^{(1)}, Z^{(2)})]$, $[\left(\left(Z^{(1)}\right)^2, \left(Z^{(2)}\right)^2\right)]$ *et* $[Z^{(1)} \cdot Z^{(2)}]$. *Cette partition sera particulièrement étudiée au chapitre 14 consacré aux surfaces de réponses.*

En conséquence on voit bien à partir de ces deux exemples qu'il faudra parler de modèle avec partition orthogonale plutôt que de modèle orthogonal. On peut parler de modèle orthogonal seulement dans la cas où, comme dans l'exemple 7.2, on considère la partition la plus fine (celle qui distingue chacun des régresseurs).

Formalisons ces exemples dans une définition.

Définition 7.5 (Orthogonalité pour des modèles réguliers) *Soit un modèle linéaire général régulier (3.5), avec*

$$Y = X \cdot \theta + \varepsilon.$$

Considérons une partition en m *termes de* X *et* θ, *soit*

$$Y = X_1 \cdot \theta_1 + \cdots + X_m \cdot \theta_m + \varepsilon,$$

où X_i *est une matrice de taille* (n, k_i) *et* $\theta_i \in \mathbb{R}^{k_i}$, *et* $k_i \in \{1, \cdots, n\}$, *pour* $i = 1, \cdots, m$ *(et avec* $\sum k_i = k$*). On dit que cette partition est orthogonale si les sous-espaces vectoriels de* $\mathbb{R}^n$,

$$[X_1], \cdots, [X_m]$$

sont orthogonaux.

Une conséquence simple de l'orthogonalité d'un modèle linéaire est que la matrice d'information $(X' \cdot X)$ a une structure bloc diagonale, chaque bloc étant associé à chaque élément de la partition.

Le plus souvent, la partition du vecteur de paramètres θ en différents effets vient

- en régression, des différentes variables ;
- en analyse de la variance, des décompositions en interactions.

L'orthogonalité donne aux modèles statistique les deux propriétés suivantes qui sont fort intéressantes :

Proposition 7.3 *Soit un modèle linéaire régulier muni d'une partition orthogonale :*

$$Y = X_1 \cdot \theta_1 + \cdots + X_m \cdot \theta_m + \varepsilon.$$

Alors :

- *les estimateurs des différents effets* $(\widehat{\theta_i})_{1\leq i\leq k}$ *sont non-corrélés (indépendants sous l'hypothèse gaussienne).*

- *pour* $\ell = 1, \cdots, m$*, l'expression de l'estimateur* $\widehat{\theta_\ell}$ *ne dépend pas de la présence ou non des autres termes* $\theta_{j'}$ *dans le modèle.*

<u>*Démonstration*</u> : L'orthogonalité implique que :

$$P_{[X]}Y = P_{[X_1]}Y + P_{[X_2]}Y + \cdots + P_{[X_m]}Y.$$

Pour $i = 1, \cdots, m$, l'estimateur $\widehat{\theta_i}$ vérifie donc

$$X_i \cdot \widehat{\theta_i} = P_{[X_i]}Y,$$

ce qui implique l'indépendance par les propriétés des lois normales isotropes. Notez que comme X est de plein rang, X_i l'est également et

$$\widehat{\theta_i} = (X_i' \cdot X_i)^{-1} \cdot X_i' \cdot Y, \tag{7.1}$$

ce qui donne la seconde assertion. ■

L'orthogonalité apporte une simplification des calculs : elle permet d'obtenir facilement une expression explicite des estimateurs. Par ailleurs, elle donne une indépendance approximative entre les tests sur les différents effets. Les tests portant sur deux effets orthogonaux ne sont liés que par l'estimation du σ^2.

Un exemple d'application de la seconde propriété est le suivant : soit un modèle de régression multiple et supposons que la partition la plus fine soit orthogonale. Alors l'expression de l'estimateur du coefficient β_l de la variable $Z^{(l)}$ vaut

$$\widehat{\beta_l} = \sum_{i=1}^{n} \frac{Z_i^{(l)} Y_i}{\left(Z_i^{(l)}\right)^2}.$$

Il suffit pour le vérifier de considérer le modèle de régression linéaire simple dans lequel on ne considère que la variable $Z^{(l)}$.

Exemple de la régression polynômiale orthogonalisée :

On considère un modèle de régression quadratique

$$Y = \mu + \beta_1 \cdot Z + \beta_2 \cdot Z^2 + \varepsilon \tag{7.2}$$

où la variable Z prend des valeurs régulièrement espacées. Sans perte de généralité on peut supposer qu'il s'agit des valeurs de de 1 à n. La moyenne empirique de Z vaut

$\overline{Z} = (n+1)/2$. On note $< .,. >$ le produit scalaire de $\mathbb{R}^n$ et on définit les variables suivantes :

$$T^{(0)} := \mathbb{1} \quad ; \quad T^{(1)} := Z - \overline{Z} \quad ; \quad T^{(2)} := (Z-\overline{Z})^2 - \frac{1}{n} < (Z-\overline{Z})^2, \mathbb{1} > .$$

En raison de la bijectivité de ce changement de variable, le modèle (7.2) est équivalent au modèle

$$Y = \gamma_0 \cdot T^{(0)} + \gamma_1 \cdot T^{(1)} + \gamma_2 \cdot T^{(2)} + \varepsilon. \tag{7.3}$$

La matrice d'information de ce nouveau modèle vaut

$$X' \cdot X = \begin{pmatrix} \|T^{(0)}\|^2 & < T^{(0)}, T^{(1)} > & < T^{(0)}, T^{(2)} > \\ < T^{(1)}, T^{(0)} > & \|T^{(1)}\|^2 & < T^{(1)}, T^{(2)} > \\ < T^{(2)}, T^{(0)} > & < T^{(2)}, T^{(1)} > & \|T^{(2)}\|^2 \end{pmatrix}.$$

Cette matrice apparaît comme la matrice de produits scalaires des vecteurs $T^{(0)}$; $T^{(1)}$; $T^{(2)}$ (appelée encore matrice de Grahm). On montre facilement que les produits scalaires $< T^{(0)}, T^{(1)} >$ et $< T^{(1)}, T^{(2)} >$ sont nuls. Par ailleurs, $T^{(2)}$ a été construit de sorte que $< T^{(2)}, \mathbb{1} >= 0$. Le modèle est donc orthogonal puisque ses régresseurs sont orthogonaux entre eux et que la matrice d'information $X' \cdot X$ est diagonale.

Comme conséquence immédiate, on obtient une formule explicite du coefficient de régression de Y sur $T^{(i)}$: il est le même que si $T^{(i)}$ était la seule variable en présence. On a donc dans le modèle (7.3),

$$\widehat{\gamma_i} = \frac{< T^{(i)}, Y >}{\|T^{(i)}\|^2}.$$

3 Orthogonalité pour des modèles non-réguliers

Définition 7.6 *Considérons la partition suivante d'un modèle linéaire*

$$Y = X_1 \cdot \theta_1 + \cdots + X_m \cdot \theta_m + \varepsilon,$$

et soit un système de contraintes $C_1 \cdot \theta_1 = 0, \cdots, C_m \cdot \theta_m = 0$ qui rendent le modèle identifiable. On dit que ces contraintes rendent la partition orthogonale si les sous-espaces vectoriels

$$V_i = \{X_i \cdot \theta_i : \theta_i \in Ker(C_i)\} \quad , \quad i = 1, \cdots, m$$

sont orthogonaux.

Cette définition prend tout son sens avec l'exemple incontournable suivant qui concerne le modèle d'analyse de la variance à deux facteurs croisés.

Proposition 7.4 *Soit le modèle d'analyse de la variance à deux facteurs croisés,*

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{i,j,k}; \quad i = 1, \cdots, I, \ j = 1, \cdots, J, \ k = 1, \cdots, n_{ij},$$

avec $n_{ij} \geq 1$, $\sum_{ij} n_{ij} = n > I \cdot J$. *Il existe des contraintes qui rendent la partition* $\mu, \alpha, \beta, \gamma$ *orthogonale si et seulement si*

$$n_{ij} = \frac{n_{i+} n_{+j}}{n_{++}} \tag{7.4}$$

où, par exemple, $n_{i+} = \sum_i n_{ij}$ *et* $n_{++} = \sum_j \sum_i n_{ij}$. *Dans ce cas, les contraintes sont*

$$(i) \ \sum_i^I \alpha_i n_{i+} = 0 \ \textit{et} \ \ (ii) \sum_j^J \beta_j n_{+j} = 0 \, ;$$

$$(iii) \ \forall i = 1, \cdots, I, \ \ \sum_j^J n_{ij} \gamma_{ij} = 0 \ \textit{et} \quad (iv) \forall j = 1, \cdots, J, \ \ \sum_i^I n_{ij} \gamma_{ij} = 0.$$

Remarques : 1/ Un énoncé très proche est possible pour le modèle additif.
2/ Notez bien que pour le système de contraintes qui correspond à la décomposition de type III que nous avons présenté au chapitre 5, qui est :

$$\sum_i \alpha_i = 0; \ \ \sum_j \beta_j = 0; \ \ \forall i, \ \sum_j \gamma_{i,j} = 0 \text{ et } \ \forall j, \ \sum_i \gamma_{i,j} = 0,$$

il n'y a possibilité d'orthogonalité (d'après la proposition 7.4) que si le modèle est **équirépété** c'est-à-dire si $n_{ij} = cte$.

Démonstration : a) Clairement, les conditions d'orthogonalité de μ et α sont équivalentes à (i), celles de μ et β sont équivalentes à (ii) , celles de l'espace engendré par μ plus α avec l'espace engendré par γ est équivalente à (iii) et enfin l'orthogonalité de l'espace engendré par μ plus β avec l'espace engendré par γ est équivalente à (iv).

b) il reste donc à examiner l'orthogonalité du sous-espace de $\mathbb{R}^n$ engendré par α et muni des contraintes ci-dessus (on le notera $[\alpha]$) avec le sous-espace engendré par β ($[\beta]$) avec le même type de contraintes. Définissons d'abord les sous-espaces suivants :

$$\begin{aligned} A &:= \{(\alpha_1, \cdots, \alpha_I) \in \mathbb{R}^I : \sum_i \alpha_i n_{i+} = 0\}, \\ \text{et symétriquement} \quad B &:= \{(\beta_1, \cdots, \beta_J) \in \mathbb{R}^J : \sum_i \beta_j n_{+j} = 0\}. \end{aligned}$$

Soit maintenant $V^{(\alpha)}$ et $V^{(\beta)}$ deux vecteurs quelconques de $[\alpha]$ et $[\beta]$, on a

$$V_{ijk}^{(\alpha)} = \alpha_i \text{ avec } \alpha = (\alpha_i)_i \in A,$$

$$V_{ijk}^{(\beta)} = \beta_j \text{ avec } \beta = (\beta_j)_j \in B.$$

Donc si (7.4) est vraie,

$$< V^{(\alpha)}, V^{(\beta)} >= \sum_{ij} n_{ij}\alpha_i\beta_j = \sum_{ij} \frac{n_{i+}\alpha_i n_{+j}\beta_j}{n_{++}} = 0.$$

c) Réciproquement, si $[\alpha]$ et $[\beta]$ sont orthogonaux, pour tout $\alpha = (\alpha_i)_i$ dans A et $\beta = (\beta_j)_j$ dans B on a

$$\sum_{ij} n_{ij}\alpha_i\beta_j = 0. \tag{7.5}$$

Fixons α. Comme la relation (7.5) est vraie pour tout β et que la seule relation vérifiée par les β_j est $\sum_j \beta_j n_{+j} = 0$, cela implique que $(\sum_i n_{ij}\alpha_i)$ est proportionnel à n_{+j} quand j varie. En sommant en j on voit que le coefficient de proportionnalité est forcément nul. On donc montré que pour tout vecteur α dans A :

$$\sum_{ij} n_{ij}\alpha_i = 0, \quad \text{pour tout } j = i, \cdots, J.$$

À nouveau cela implique que le vecteur n_{ij} est proportionnel a n_{i+}. Notons C_j le coefficient de proportionnalité, soit $n_{ij} = C_j n_{i+}$, et sommons en j ; on obtient alors

$$C_j n_{++} = n_{+j}.$$

Il suffit de re-injecter cette formule dans la précédente pour obtenir le résultat. ■

4 Exercices

Exercice 7.1

(*) Montrer l'équivalence (avec nos notations maintenant classiques) :

$$X \text{ de plein rang} \Leftrightarrow X' \cdot X \quad \text{inversible.}$$

Exercice 7.2

(*) Quel est l'énoncé correspondant à la proposition 7.4 pour le modèle additif ? Même question pour le modèle hiérarchisé.

Chapitre 8

Propriétés asymptotiques

Ce chapitre étudie les comportements asymptotiques des différents estimateurs et statistiques de test intervenant dans un modèle linéaire, et met en évidence le fait que les tests de Student et de Fisher peuvent être utilisés asymptotiquement même si le modèle n'est pas gaussien.

1 Introduction

1.1 Qu'appelle-t-on asymptotique ?

Le modèle linéaire gaussien se caractérise par un grand nombre de propriétés dites "à distance finie", c'est-à-dire des propriétés valides pour un nombre de données n qui n'a pas besoin d'être grand. Même dans ce cadre très confortable, certaines questions asymptotiques se posent, comme par exemple celle de la convergence des estimateurs quand $n \to +\infty$. Dans le cadre non-gaussien, la question de l'étude asymptotique est plus cruciale car on ne connaît pas la loi des estimateurs. Il devient donc nécessaire de travailler avec un grand nombre de données pour pouvoir utiliser les convergences des différentes statistiques présentées. Alors et alors seulement, on pourra construire des intervalles de confiance ainsi que des tests.

De manière générale, trois options se dégagent des résultats :

– Les résultats asymptotiques peuvent être identiques que l'on soit dans le cadre d'un modèle linéaire gaussien ou d'un modèle linéaire non gaussien. Nous traiterons alors les deux cas simultanément. C'est le cas le plus fréquent

– D'autres résultats asymptotiques peuvent être également identiques dans les cas gaussiens et non gaussiens lorsque l'on aura adjoint aux distributions non gaussiennes certaines conditions (cela concerne par exemple la distribution de $\widehat{\theta}$ et celle de $X \cdot \widehat{\theta}$, pour lesquelles on obtient la convergence vers une distribution normale sous des conditions relativement faibles).

– Enfin, certains résultats asymptotiques seront dépendant des distributions initiales et différeront clairement entre modèle gaussien et modèle non-gaussien (ce sera notamment le cas de la vitesse de convergence de $\widehat{\sigma^2}$ qui dépend de manière générale du kurtosis , relatif aux moments d'ordre 2 et 4 de la loi). On ne pourra obtenir de "bons" résultats asymptotiques que pour des modèles possédant un grand nombre de degrés de liberté résiduels.

D'un point de vue statistique, et notamment dans le cadre de la statistique asymptotique, on peut différencier deux modes distincts de convergence : d'une part, la notion de convergence en loi (notée $\xrightarrow[n\to+\infty]{\mathcal{L}}$), et d'autre part, les convergences en probabilité (notée $\xrightarrow[n\to+\infty]{\mathcal{P}}$), en moyenne quadratique (notée $\xrightarrow[n\to+\infty]{m.q.}$) et presque-sûre (notée $\xrightarrow[n\to+\infty]{p.s.}$). Les différences entre ces trois dernières convergences sont des finesses théoriques, et de manière générale elles impliquent toutes les trois sous des formes diverses la convergence des réalisations. Quant à la convergence en loi, elle se différencie facilement des 3 autres, car comme son nom l'indique, elle ne concerne que la loi de probabilité.

Pour obtenir des comportements asymptotiques, nos outils seront, classiquement, la loi des grands nombres et le théorème de la limite centrale. Nous renvoyons à l'annexe pour des précisions sur ces théorèmes dans le cas général. Cependant, pour aller un peu plus loin, nous utiliserons également une version plus forte du théorème de la limite centrale (on trouvera plus de détails dans, par exemple, Feller [26]) :

Théorème 8.1 (Théorème de la limite centrale de Lindeberg) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires indépendantes et de même loi non gaussienne admettant un moment d'ordre* 2. *Sans perte de généralité, on peut supposer que* $\mathbb{E}(X_i) = 0$, $Var(X_i) = 1$ *pour tout* $i \in \mathbb{N}$. *Considérons le "tableau triangulaire" suivant :*

$$(a_i^n)_{n\in\mathbb{N},i=1,\cdots,n} \quad \textit{tel que} \quad \sum_{i=1}^{n}(a_i^n)^2 = 1 \quad \textit{pour tout } n \in \mathbb{N}^*.$$

Alors on a :

$$Z_n = \sum_{i=1}^{n} a_i^n X_i \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1) \tag{8.1}$$

si et seulement si (en abrégé désormais ssi)

$$\max_{1\le i\le n} |a_i^n| \xrightarrow[n\to+\infty]{} 0. \tag{8.2}$$

Remarque : Dans l'énoncé de ce théorème, on précise la non-gaussianité de la loi des X_i. Si les X_i sont des variables aléatoires gaussiennes, il est bien clair que sous les hypothèse du théorème, la convergence (8.1) est vraie. En réalité il y a égalité pour tout n (et non seulement quand $n \to \infty$).

1.2 Hypothèses et notations

Dans la suite de ce chapitre nous considérerons, sauf expressément précisé, une famille de modèles linéaires

$$Y^n = X^n \cdot \theta^n + \varepsilon^n \tag{8.3}$$

où $Y^n = (Y_i^n)_{1\le i\le n}$ est un vecteur connu de taille n, X^n est une matrice connue d'ordre (n, k_n), $k_n < n$ (sans perte de généralité, **on peut supposer X^n de plein rang k_n**), θ^n est un vecteur inconnu de taille k_n et le vecteur $\varepsilon^n = (\varepsilon_i^n)_{1\le i\le n}$ est un vecteur aléatoire inconnu de taille n constitué de n variables centrées, indépendantes et identiquement distribuées (dont la loi commune ne dépend pas de n) et de variance finie.

Remarque : Pour ce modèle, nous noterons **H** l'ensemble des hypothèses faites sur les erreurs ε_i et qui comprennent : les postulats **P1-3**, plus l'équidistribution des erreurs. En revanche, le postulat **P4** de gaussianité des erreurs ne sera supposé que dans quelques cas expressément mentionnés.

Par la suite, pour alléger les notations, nous noterons encore Y, X, θ et ε au lieu de Y^n, X^n, θ^n et ε^n. En revanche, les estimateurs, comme par exemple $\widehat{\theta}^n$ ou $\widehat{\sigma}^{2,n}$, seront toujours notés avec le n.

Le cadre de l'étude sera différent suivant que la taille k_n de θ dépend de n ou non (on notera alors $k_n = k$). Donnons un exemple de chaque situation :

– On considère le modèle du plan en bloc complets (voir chapitre 11) dans lequel on pose un modèle d'analyse de la variance à deux facteurs additifs avec t traitements et b blocs de longueur t. Le nombre de traitements reste fixe, mais le

nombre de blocs tend vers l'infini. On a alors $n = b \cdot t$ et la dimension k_n du modèle vaut $b + t - 1$ qui tend vers l'infini également.

– Dans un modèle de régression multiple avec 5 régresseurs, on augmente le nombre des observations. Dans ce cas n tend vers l'infini mais en revanche k reste fixe, égal à 6 (cinq régresseurs plus le terme constant).

On définit H^n_{ij} comme étant l'élément (i, j) de la matrice, connue, du projecteur sur le sous-espace vectoriel engendré par X, $H^n = P_{[X]}$. Cette matrice, qui sera essentiellement la clé de voûte du comportement asymptotique du modèle linéaire, est appelée parfois "hat matrix" car elle associe $\widehat{Y}$ à Y et se détermine par :

$$H^n = X \cdot (X' \cdot X)^{-1} \cdot X'.$$

(rappelons que l'on a supposé X régulière. Dans le cas général, la détermination de la matrice $P_{[X]}$ nécessite l'utilisation d'inverses généralisées, voir chapitre 7). On définit également la norme matricielle

$$\|H^n\| = \max_{1 \leq i \leq n} |H^n_{ii}|.$$

2 Comportement asymptotique des statistiques

Revenons donc maintenant sur différentes statistiques définies au cours des chapitres précédents et plus particulièrement au cours du chapitre 3. En premier lieu, examinons si les estimateurs des paramètres sont convergents lorsque le nombre de données (c'est-à-dire n) devient grand. Dans un deuxième temps, nous étudierons ce qu'il advient des statistiques de test, $\widehat{F}^n$ ou $\widehat{T}^n$ lorsque, pareillement, n devient grand.

2.1 Comportement asymptotique des estimateurs

Soit la famille de modèles linéaires (8.3). En prenant le cas particulier de la régression simple, on conçoit bien que, sous certaines conditions, plus l'on dispose de données, plus la droite de régression va "se caler" sur la "vraie" droite d'équation $y = \mu + \beta \cdot x$: le vecteur $\widehat{\theta}^n = (\widehat{\beta}^n, \widehat{\mu}^n)'$ estimateur, qui définit entièrement cette droite, va converger vers la "vraie" valeur θ. Et il peut être particulièrement important de préciser l'erreur faite sur cette estimation ; on pense notamment à des expériences en physique dans lesquelles l'erreur ε^n joue le rôle d'erreur de mesure, ou bien des mesures de la volatilité (représentée par σ^2) pour des données financières. La conséquence d'une "bonne" estimation des paramètres est de permettre d'obtenir des prédictions avec une marge d'erreur connue. Revenons donc au cas général.

On estime le vecteur des paramètres θ^n (de taille k_n) par moindres carrés. On a ainsi, comme vu dans les chapitres précédents,

$$\widehat{\theta}^n = (X' \cdot X)^{-1} \cdot X' \cdot Y \quad \text{et} \quad \widehat{Y}^n = H^n \cdot Y = X \cdot \widehat{\theta}^n.$$

avec $\widehat{Y}^n = (\widehat{Y}_i^n)_{1 \leq i \leq n}$. Notons également $X \cdot \theta = \big((X \cdot \theta)_i\big)_{1 \leq i \leq n}$. On peut alors obtenir la proposition suivante :

Proposition 8.1 *Dans le cadre de la famille de modèles linéaires (8.3) munis de l'hypothèse* **H** *on a :*

1/ Pour i fixé dans $\{1, \cdots, n\}$,
$\left(\widehat{Y}_i^n - (X \cdot \theta)_i\right) \underset{n \to +\infty}{\overset{m.q.}{\longrightarrow}} 0$ *ssi* $H_{ii}^n \underset{n \to +\infty}{\longrightarrow} 0$.

2/ Pour tout $i \in \{1, \cdots, n\}$,
$\left(\widehat{Y}_i - (X \cdot \theta)_i\right) \underset{n \to +\infty}{\overset{m.q.}{\longrightarrow}} 0$ *ssi* $\|H^n\| = \max\limits_{1 \leq i \leq n} |H_{ii}^n| \underset{n \to +\infty}{\longrightarrow} 0$.

Ces convergences en moyenne quadratique entraînent les convergences en probabilité.

<u>*Démonstration*</u> : 1/ Supposons que $H_{ii}^n \underset{n \to +\infty}{\longrightarrow} 0$. On sait que $\widehat{Y}_i^n$ est un estimateur sans biais de Y_i, il suffit donc de calculer sa variance. Il est facile de voir que

$$\mathrm{Var}\,(\widehat{Y}_i^n) = \mathrm{Var}\left(\sum_{j=1}^n H_{ij}^n \cdot Y_j\right) = \sigma^2 \cdot \sum_{j=1}^n (H_{ij}^n)^2 = H_{ii}^n \cdot \sigma^2,$$

car H^n est un projecteur orthogonal ce qui implique que $(H^n)' \cdot H^n = H^n$. On obtient ainsi l'équivalence en ce qui concerne la convergence en moyenne quadratique.

2/ Une fois obtenue la première partie de la preuve, il est clair que l'hypothèse $\|H^n\| \underset{n \to +\infty}{\longrightarrow} 0$ étant équivalente au fait que pour tout $i \in \{1, \cdots, n\}$, $H_{ii}^n \underset{n \to +\infty}{\longrightarrow} 0$, elle est également équivalente à $\left(\widehat{Y}_i^n - (X \cdot \theta)_i\right) \underset{n \to +\infty}{\overset{m.q.}{\longrightarrow}} 0$ pour tout $i \in \{1, \cdots, n\}$. ■

Remarquons que si H_{ii}^n ne tend pas vers zéro quand $n \to \infty$, il est montré dans Huber [33] (page 157) qu'il ne peut pas y avoir convergence en probabilité.

Remarquons également que l'on on ne peut pas espérer obtenir la convergence en moyenne quadratique du vecteur $\widehat{Y}^n$ vers le vecteur Y. En effet, le risque quadratique

de cette estimation, qui est défini par $R^n = \mathbb{E}\,||\widehat{Y}^n - X\cdot\theta||_n^2$, où $||.||_n$ désigne la norme euclidienne classique sur $\mathbb{R}^n$, vérifie :

$$\begin{aligned} R^n &= \sum_{i=1}^n \mathbb{E}\left(\widehat{Y}_i^n - \left(X\cdot\theta\right)_i\right)^2 \\ &= \sigma^2\cdot\sum_{i=1}^n H_{ii}^n = \sigma^2\cdot\mathrm{Tr}(H^n) \\ &= \sigma^2\cdot k_n \end{aligned}$$

qui ne tend pas vers 0 (et peut même tendre vers $+\infty$ lorsque $k_n\to\infty$). Nous en resterons donc là quant à la convergence de l'estimateur $\widehat{Y}^n$.

En revanche, on peut obtenir des résultats quant à la convergence asymptotique de l'estimateur de la variance de l'erreur, $\widehat{\sigma}^{2,n} = \dfrac{1}{n-k_n}\displaystyle\sum_{i=1}^n\left(\widehat{Y}_i^n - Y_i\right)^2$:

Proposition 8.2 *Dans le cadre de la famille de modèles linéaires (8.3) munis de l'hypothèse* **H**,

1/ Si le modèle est gaussien (postulat **P4***), et si* $\lim_{n\to\infty}(n-k_n) = +\infty$,

$$\widehat{\sigma}^{2,n} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \sigma^2 \quad ; \quad \sqrt{n-k_n}(\widehat{\sigma}^{2,n}-\sigma^2) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,2\sigma^4).$$

2/ Quelque soit la loi des erreurs ε_i *du modèle, sous l'hypothèse* $\lim_{n\to\infty}\dfrac{k_n}{n} = 0$, *alors* $\widehat{\sigma}^{2,n} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \sigma^2$.

3/ Quelque soit la loi des erreurs ε_i *du modèle, si cette loi admet un moment d'ordre 4, en notant* $\mu_4 = \mathbb{E}\,(\varepsilon_i^4)$, *et sous l'hypothèse* $\lim_{n\to\infty}\dfrac{k_n}{\sqrt{n}} = 0$, *on obtient :*

$$\sqrt{n}(\widehat{\sigma}^{2,n}-\sigma^2) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0,\mu_4-\sigma^4).$$

<u>*Démonstration*</u> : 1/ Sous l'hypothèse gaussienne, on sait que pour tout n tels que $n-k_n\geq 1$,

$$\widehat{\sigma}^{2,n} \overset{\mathcal{L}}{\sim} \sigma^2\cdot\frac{1}{n-k_n}\cdot\chi^2(n-k_n).$$

La convergence découle directement du Théorème de la Limite Centrale appliqué à la loi du $\chi^2(n-k_n)$, qui rappelons-le, se comporte comme la somme de $(n-k_n)$ carrés

de variables gaussiennes centrées réduites indépendantes.

2/ Si le modèle n'est pas gaussien, on remarque que sous l'hypothèse $\lim_{n\to\infty} \frac{k_n}{n} = 0$, $\widehat{\sigma}^{2,n}$ est asymptotiquement équivalent en loi à l'estimateur :

$$\tilde{\sigma}^{2,n} = \frac{1}{n} \|P_{[X^\perp]} \cdot Y\|^2 = \frac{1}{n} \|P_{[X^\perp]} \cdot \varepsilon\|^2.$$

Du fait de la propriété de non-biais de l'estimateur $\widehat{\sigma}^{2,n}$ on a

$$\mathbb{E}\,(\tilde{\sigma}^{2,n}) \underset{n\to+\infty}{\longrightarrow} \sigma^2.$$

Par ailleurs, si on définit

$$V^n := \frac{1}{n} \|\varepsilon\|^2,$$

on a $\mathbb{E}\,(V^n) = \sigma^2$ et par la Loi Faible des Grands Nombres, du fait que $\|\varepsilon\|^2 = \varepsilon_1^2 + \cdots + \varepsilon_n^2$, on a

$$V^n \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \sigma^2.$$

Pour finir, nous avons $V^n - \tilde{\sigma}^{2,n} \geq 0$ (car la norme du projeté d'un vecteur est inférieure à celle du vecteur) et $\mathbb{E}\left(V^n - \tilde{\sigma}^{2,n}\right) \underset{n\to+\infty}{\longrightarrow} 0$. Ceci entraîne que $V^n - \tilde{\sigma}^{2,n} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} 0$, car une variable positive dont l'espérance tend vers 0 tend en probabilité vers 0 (on utilise l'inégalité de Markov). Par conséquent, $\tilde{\sigma}^{2,n}$ et donc $\widehat{\sigma}^{2,n}$ héritent de la propriété de convergence en probabilité de V^n vers σ^2 .

3/ Pour montrer ce théorème de la limite centrale pour $\widehat{\sigma}^{2,n}$, on va reprendre les arguments de la preuve précédente en les reprécisant. On peut ainsi écrire que

$$\widehat{\sigma}^{2,n} = \frac{1}{n-k_n} \cdot \varepsilon' \cdot H^n \cdot \varepsilon.$$

(on a utilisé le fait que $(H^n)' \cdot H^n = H^n$). Mais, comme H^n est un projecteur sur un sous-espace de dimension $(n-k_n)$, on peut écrire qu'il existe P^n matrice orthogonale telle que $H^n = (P^n)' \cdot I(n-k_n) \cdot P^n$, où $I(n-k_n)$ est la matrice diagonale avec $(n-k_n)$ uns et k_n zéros sur la diagonale. Alors :

$$\begin{aligned}
\widehat{\sigma}^{2,n} &= \frac{1}{n-k_n} \varepsilon' \cdot (P^n)' \cdot I(n-k_n) \cdot P^n \cdot \varepsilon \\
&= \frac{1}{n-k_n} \sum_{i=k_n+1}^{n} e_i^2 \\
&= \frac{1}{n-k_n} \sum_{i=1}^{n} \varepsilon_i^2 - \frac{1}{n-k_n} \sum_{i=1}^{k_n} e_i^2, \qquad (8.4)
\end{aligned}$$

avec $e = (e_i)_{1\leq i\leq n} = P^n \cdot \varepsilon$. Pour tout $i \in \{1, \cdots, n\}$, on a $\mathbb{E}(e_i) = 0$ et $\mathbb{E}(e_i \cdot e_j) = \sigma^2 \cdot \delta_{ij}$ avec $j \in \{1, \cdots, n\}$, du fait que $(P^n)' \cdot P^n = I$. Par suite, sous l'hypothèse $\lim_{n\to\infty} \frac{k_n}{\sqrt{n}} = 0$

$$\sqrt{n}\frac{1}{n-k_n}\sum_{i=1}^{k_n} e_i^2 \xrightarrow[n\to+\infty]{\mathcal{P}} 0 \tag{8.5}$$

car $\sqrt{n}\frac{1}{n-k_n}\sum_{i=1}^{k_n} e_i^2 \geq 0$ et $\mathbb{E}\left(\sqrt{n}\frac{1}{n-k_n}\sum_{i=1}^{k_n} e_i^2\right) = \sigma^2 \cdot \frac{\sqrt{n}\cdot k_n}{n-k_n} \xrightarrow[n\to+\infty]{} 0$ (on utilise encore l'inégalité de Markov). De plus,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \sigma^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \mu_4 - \sigma^4), \tag{8.6}$$

d'après le Théorème de la Limite Centrale appliqué à la suite $(\varepsilon_i^2)_i$, suite de variables aléatoires indépendantes et identiquement distribuées possédant un moment d'ordre 2 et telles que $\mathrm{Var}(\varepsilon_i^2) = \mu_4 - \sigma^4$. Donc, comme $\lim_{n\to\infty} \frac{k_n}{\sqrt{n}} = 0$, on a :

$$\sqrt{n}\left(\frac{1}{n-k_n}\sum_{i=1}^{n}\varepsilon_i^2 - \sigma^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \mu_4 - \sigma^4).$$

Par suite, en reprenant (8.4), $\sqrt{n}(\widehat{\sigma}^{2,n} - \sigma^2) = X_n - Y_n$ avec $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \mu_4 - \sigma^4)$ d'après (8.6) et $Y_n \xrightarrow[n\to+\infty]{\mathcal{P}} 0$. Ainsi $\sqrt{n}(\widehat{\sigma}^{2,n} - \sigma^2)$ à la même convergence en loi que X_n, ce qui implique que :

$$\sqrt{n}(\widehat{\sigma}^{2,n} - \sigma^2) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \mu_4 - \sigma^4). \qquad \blacksquare$$

Remarquons que dans le cas gaussien $\mu_4 = 3\sigma^4$, on retrouve par 3/ le théorème de la limite centrale 1/. Cependant, en général, pour les autres lois, la vitesse de convergence sera différente.

Pour établir la normalité asymptotique $\widehat{\theta}^n$, nous allons procéder de la façon la plus générale possible. Pour commencer, définissons ce que l'on appellera normalité asymptotique :

Définition 8.1 *Nous dirons qu'une suite de vecteurs aléatoires Z^n dont la taille peut dépendre de n est asymptotiquement gaussienne si pour toute suite de combinaisons linéaires $(C^n)' \cdot Z^n$ non nulles, on a*

$$U^n := \frac{(C^n)' \cdot Z^n - \mathbb{E}((C^n)' \cdot Z^n)}{\sqrt{Var((C^n)' \cdot Z^n)}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Nous pouvons maintenant étudier la normalité asymptotique de $\widehat{\theta}^n$. Nous commençons par une première proposition, évidente, mais qui a la vertu de situer le problème :

Proposition 8.3 *On considère la famille de modèles linéaires (8.3) sous l'hypothèse* **H** *et supposons la dimension k_n fixée. Alors :*

$$(X' \cdot X)^{-1} \underset{n\to+\infty}{\longrightarrow} 0 \iff \widehat{\theta}^n \underset{n\to+\infty}{\overset{m.q.}{\longrightarrow}} \theta.$$

Démonstration : Sous les hypothèses de la proposition, on a la formule **F3** concernant la matrice de variance-covariance de $\widehat{\theta}^n$, soit $\mathrm{Var}\,(\widehat{\theta}^n) = \sigma^2 \cdot (X' \cdot X)^{-1}$. L'absence de biais permet alors de conclure. ■

Les deux résultats suivants précisent le comportement et la normalité asymptotique des estimateurs $\widehat{Y^n}$ et $\widehat{\theta}^n$ dans le cas général.

Théorème 8.2 *On considère la famille de modèles linéaires (8.3) sous l'hypothèse* **H**. *Alors, sous l'hypothèse $\|H^n\| = \max\limits_{1\le i\le n} |H^n_{ii}| \underset{n\to+\infty}{\longrightarrow} 0$, les estimateurs $\widehat{Y}^n$ et $\widehat{\theta}^n$ sont asymptotiquement gaussiens.*
Réciproquement si la loi des ε_i n'est pas gaussienne et si la condition $\|H^n\| = \max\limits_{1\le i\le n} |H^n_{ii}| \underset{n\to+\infty}{\longrightarrow} 0$ n'est pas vérifiée, alors $\widehat{Y}^n$ et $\widehat{\theta}^n$ ne sont pas asymptotiquement gaussiens.

Remarque : La condition $\|H^n\| \underset{n\to+\infty}{\longrightarrow} 0$ apparaît donc en pratique comme un condition nécessaire et suffisante à la fois pour la convergence et pour la normalité asymptotique.

Démonstration : Nous donnons la démonstration seulement de la première partie. Pour ce qui est de la réciproque, qui a moins d'intérêt en pratique, nous renvoyons à Huber [33], dont il faut adapter les arguments car notre résultat est un petit peu plus général.

On suppose donc que $\|H^n\| \underset{n\to+\infty}{\longrightarrow} 0$. Nous allons montrer la normalité asymptotique de $\widehat{Y}^n = H^n \cdot Y$ (celle de $\widehat{\theta}^n$ en découlera puisque $\widehat{Y}^n = X \cdot \widehat{\theta}^n$). En utilisant le fait déjà vu que $\mathbb{E}\,\widehat{\theta}^n = \theta$, on peut écrire que :

$$U^n = \frac{(C^n)' \cdot H^n \cdot Y - (C^n)' \cdot H^n \cdot X \cdot \theta}{\sqrt{(C^n)' \cdot \mathrm{Var}\,(\widehat{Y}_n) \cdot C^n}} = \frac{(C^n)' \cdot H^n \cdot \varepsilon}{\sigma \cdot \sqrt{(C^n)' \cdot H^n \cdot C^n}} = (V^n)' \cdot \tilde{\varepsilon},$$

avec :

$$(V^n)' = \frac{(C^n)' \cdot H^n}{\sqrt{(C^n)' \cdot H^n \cdot C^n}} \quad \text{et} \quad \tilde{\varepsilon} = \frac{\varepsilon}{\sigma}.$$

On a donc une suite de vecteurs bien normalisée : $\|(V^n)'\| = 1$. Le théorème central limite de Lindeberg 8.1 s'applique donc si

$$\max_{1\leq i\leq n} V_i^n \underset{n\to+\infty}{\longrightarrow} 0.$$

On écrit $V^n := B^n \cdot b^n$ avec

$$\begin{aligned} B^n &:= (H^n)^{1/2}, \text{ matrice de taille } (n, k_n) \text{ et} \\ b^n &:= \frac{(H^n)^{1/2} \cdot C^n}{\sqrt{(C^n)' \cdot H^n \cdot C^n}} \text{ vecteur de taille } k_n, \end{aligned}$$

où $(H^n)^{1/2}$ est une matrice vérifiant $((H^n)^{1/2})' \cdot (H^n)^{1/2} = H^n$.

$$\begin{aligned} \text{Or } (V_i^n)^2 = (B^n \cdot b^n)_i^2 &= \big(\sum_{j=k} B_{ij}^n b_{n,j}\big)^2 \\ &\leq \big(\sum_{j=1}^{k} (B_{ij}^n)^2\big) \cdot \big(\sum_{j=1}^{k} (b_j^n)^2\big) \text{ (Inégalité de Cauchy-Schwarz),} \\ &\leq \sum_{j=1}^{k} (B_{ij}^n)^2 \text{ car } (b^n)' \cdot b^n = \sum_{j=1}^{k} (b_j^n)^2 = 1 \\ &\leq H_{ii}^n \text{ puisque } (B^n)' \cdot B^n = H^n. \end{aligned}$$

On en déduit donc la convergence de V_i^n vers 0 sous l'hypothèse $\|H^n\| \underset{n\to+\infty}{\longrightarrow} 0$, et donc la normalité asymptotique de $\widehat{Y}^n$ sous cette condition. ■

On peut donner un cas particulier intéressant découlant en partie de cette proposition :

Corollaire 8.1 *On considère la famille de modèles linéaires (8.3) munis de l'hypothèse* **H**. *Supposons de plus que la dimension k est fixée.*

i. Si $\frac{1}{n^\alpha}(X'^n \cdot X^n) \underset{n\to+\infty}{\longrightarrow} Q$, *où* $\alpha > 0$ *et* Q *est une matrice d'ordre k définie positive, et si* $\|H^n\| = \max_{1\leq i\leq n} |H_{ii}^n| \underset{n\to+\infty}{\longrightarrow} 0$, *alors :*

$$n^{\alpha/2}(\widehat{\theta}^n - \theta) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, \sigma^2 \cdot Q^{-1}). \tag{8.7}$$

ii. La relation (8.7) est également vraie sous la seule hypothèse $\frac{1}{n^\alpha}(X'^n \cdot X^n) \underset{n\to+\infty}{\longrightarrow} Q$ *dans le cadre du modèle linéaire gaussien.*

Démonstration : La preuve de i. est une simple application de la Proposition 8.2. La preuve de ii. est immédiate du fait de la gaussianité de $\widehat{\theta}^n$. ■

2.2 Comportement asymptotique des statistiques de test

On se place toujours dans le cadre de la famille de modèles linéaires statistiques (8.3) muni de l'hypothèse **H**. Que se passe-t-il pour les différentes statistiques de test lorsque les effectifs deviennent grands ?
Pour répondre à cette question, on commence par revenir au cas gaussien qui demande d'étudier le comportement asymptotique des lois de Fisher et de Student :

Proposition 8.4 *Soit p un entier fixé tel que $0 < p$. Alors :*

1/ si $(\widehat{F}^n)_{n>0}$ est une suite de variables aléatoires de loi de Fisher de paramètres (p, n), alors

$$\widehat{F}^n \xrightarrow[n\to+\infty]{\mathcal{L}} \frac{1}{p} \cdot \chi^2(p).$$

2/ si $(\widehat{T}^n)_{n>0}$ est une suite de variables aléatoires de loi de Student de paramètre n, alors

$$\widehat{T}^n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Démonstration : pour montrer 1/, il suffit d'écrire pour $n > p$, $\widehat{F}^n$ comme le rapport de deux variables aléatoires indépendantes qui suivent respectivement une loi $\frac{1}{p} \cdot \chi^2(p)$ (numérateur) et une loi $\frac{1}{n} \cdot \chi^2(n)$ (dénominateur). Or, lorsque $n \to \infty$, ce dénominateur tend en probabilité vers 1 (voir la démonstration de la proposition 8.2). Par suite, on a un produit de variables dont une converge en probabilité vers 1 : ce produit converge en loi vers la loi du numérateur.

Pour montrer 2/, il suffit là-encore d'écrire $\widehat{T}^n$ comme le rapport de deux variables aléatoires indépendantes qui suivent une loi $\mathcal{N}(0,1)$ au numérateur et une loi $\sqrt{\frac{1}{n} \cdot \chi^2(n)}$ au dénominateur. Le dénominateur tend en probabilité vers 1 quand $n \to \infty$, d'où la convergence en loi de $\widehat{T}^n$ vers la loi $\mathcal{N}(0,1)$. ■

Ces résultats asymptotiques peuvent donc s'appliquer aux tests de Fisher et de Student du chapitre 3 avec k fixé, dès que les postulats **P1-4** sont respectés. Que se passe-t-il maintenant lorsque la famille de modèles linéaires (8.3) n'est plus gaussienne, c'est-à-dire que le postulat **P4** n'est plus supposé, mais que l'on a tout de même l'hypothèse **H** ? Nous répondrons à cette question en ne considérant que le comportement des

statistiques de Fisher $\widehat{F}$ (voir proposition 3.3), le cas des statistiques de Student $\widehat{T}$ n'étant qu'un cas particulier (avec $p = 1$). En combinant les résultats précédents de convergence des estimateurs de σ^2 et de θ, et en utilisant les outils du Théorème 8.2, on peut obtenir le résultat suivant :

Théorème 8.3 *On considère la famille de modèles linéaires (8.3) munis de l'hypothèse* **H**. *On effectue la suite de tests de Fisher définis pour $n \in \mathbb{N}^*$ par les hypothèses nulles $\mathcal{H}_0^n$: $(C^n)' \cdot \theta^n = 0$ et les l'hypothèses alternatives $\mathcal{H}_1^n$: $(C^n)' \cdot \theta^n \neq 0$, avec $(C^n)_{n\in\mathbb{N}^*}$ une suite de matrices d'ordre (k_n, p) et de rang $p \leq k_n$, où p est un entier fixé. On considère la suite de statistiques de test $(\widehat{F}^n)_{n>k_n}$. On a*

$$\widehat{F}^n = \frac{\widehat{\theta}^n \cdot C^n \cdot ((C^n)' \cdot (X' \cdot X)^{-1} \cdot C^n)^{-1} \cdot (C^n)' \cdot \widehat{\theta}^n}{p \cdot \widehat{\sigma}^{2,n}}.$$

Alors, sous la suite d'hypothèses nulles $(\mathcal{H}_0^n)_n$,

$$si \lim_{n\to\infty} \frac{k_n}{n} = 0 \quad et \; si \quad \|H^n\| = \max_{1\leq i\leq n} |H_{ii}^n| \underset{n\to+\infty}{\longrightarrow} 0, \quad \widehat{F}^n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \frac{1}{p} \cdot \chi^2(p).$$

<u>*Démonstration*</u> : On utilise les résultats de convergence précédents concernant $\widehat{\theta}^n$ (asymptotiquement gaussien) et $\widehat{\sigma}^{2,n}$ (convergeant en probabilité vers σ^2). Le produit des différents comportements asymptotiques amène au résultat. ■

Remarques : En pratique on constate par simulation que la limite ci-dessus est atteinte le plus souvent très tôt c'est à dire avec des effectifs de l'ordre de quelques dizaines.

Cette loi limite $\frac{1}{p} \cdot \chi^2(p)$ est la même que celle obtenue dans le cas gaussien, la différence étant que dans le cas gaussien, la matrice H^n ne doit rien vérifier et qu'il suffit seulement d'avoir $n - k_n \to \infty$. Le théorème 8.3 valide en fait le test du χ^2 comme limite du test de Fisher quand le nombre de degrés de liberté résiduel tend vers l'infini et sous certaines conditions.

2.3 Commentaires sur les hypothèses faites sur le modèle

Au cours des différents résultats asymptotiques précédents, nous avons été amenés à émettre plusieurs hypothèses sur la famille de modèles linéaires (8.3). En plus des postulats **P1-4** ou de l'hypothèse **H** nous avons rencontré les hypothèses :

- Hypothèse (A1) : $\|H^n\| = \max_{1\leq i\leq n} |H_{ii}^n| \underset{n\to+\infty}{\longrightarrow} 0$;
- Hypothèse (A2) : il existe $\alpha > 0$ et Q une matrice définie positive tels que $\frac{1}{n^\alpha} (X' \cdot X) = Q$;

- Hypothèse (A3) : $(X' \cdot X)^{-1} \underset{n\to+\infty}{\longrightarrow} 0$;
- Hypothèse (B1) : $\lim_{n\to\infty} \frac{k_n}{n} = 0$ (la taille k_n de θ^n est négligeable devant n) ;
- Hypothèse (B2) : $\lim_{n\to\infty} \frac{k_n}{\sqrt{n}} = 0$ (la taille k_n de θ^n est négligeable devant $\sqrt{n}$) ;
- Hypothèse (B3) : $\lim_{n\to\infty}(n - k_n) = +\infty$ (le nombre de degrés de liberté résiduel $(n - k_n)$ tend vers l'infini) ;

Il est clair que (B2) $\Longrightarrow$ (B1) $\Longrightarrow$ (B3), et que les réciproques sont fausses. Dans les applications, il est facile de savoir lorsqu'une de ces conditions est vérifiée, et on peut dire que la condition (B3) est peu coûteuse (mais elle ne s'applique qu'au cas gaussien). Remarquons que dans le cas où la taille k_n du modèle est constante, les trois hypothèses (B1) (B2) et (B3) sont vérifiées.

En revanche, les hypothèses (A1-3) sont plus délicates à vérifier expérimentalement. En fait, et nous en verrons quelques exemples ci-dessous, se greffent derrière la vérification de ces hypothèses la question du choix de la modélisation.

Pour commencer, d'une manière générale, si on peut dire de façon évidente que l'hypothèse (A2) entraîne l'hypothèse (A3), il n'y a pas de relation d'implication simple entre (A1) et (A2) ou (A3). Prenons par exemple le cas simple où $X = (x_i)_{1\leq i\leq n}$. Alors (A1) est équivalente à $\max_{i\in\{1,\cdots,n\}} \left(\frac{x_i^2}{\sum_{i=1}^n x_i^2}\right) \underset{n\to+\infty}{\longrightarrow} 0$ et (A3) est équivalente à $\sum_{i=1}^n x_i^2 \underset{n\to+\infty}{\longrightarrow} +\infty$. Si $x_1 = n$ et $x_i = 0$ pour $i \geq 2$, alors on a (A3) mais on n'a pas (A1). Si maintenant $x_i = n^{-1/2}$ pour tout $i \in \mathbb{N}^*$, alors on a (A1) mais on n'a pas (A3).

Malgré tout, on peut dire qu'il existe des cas simples et suffisamment génériques où les hypothèses $(X' \cdot X)^{-1} \underset{n\to+\infty}{\longrightarrow} 0$ et $\frac{1}{n}(X' \cdot X) \underset{n\to+\infty}{\longrightarrow} Q$ sont vérifiées. Par exemple, dans le cadre de l'analyse de la variance à 1 facteur sur k classes, la matrice $X' \cdot X$ est une matrice diagonale d'ordre k, dont les termes diagonaux sont les effectifs de chaque classes (la somme de tous ces effectifs valant n). Dès que la fréquence de chaque classe tend vers une constante, alors on a bien $\frac{1}{n}(X' \cdot X) \underset{n\to+\infty}{\longrightarrow} Q$ (c'est notamment le cas lorsque les classes ont toutes le même effectif). Un autre exemple parlant est celui de la régression. Lorsque chaque colonne j de X est constitué de réalisations indépendantes d'une variable aléatoire Z^j de carrés intégrable, les différentes Z^j étant indépendantes

entre elles, alors la Loi des Grands Nombres montre que $\frac{1}{n}(X' \cdot X) \underset{n\to+\infty}{\longrightarrow} D$, où D est une matrice diagonale dont les termes diagonaux sont les $\mathbb{E}\left((Z^j)^2\right)$.

Cependant, même dans des cas simples (et souvent rencontrés), les conditions demandées sur $(X' \cdot X)$ ou sur $\|H^n\|$ ne sont pas toujours respectées. A titre d'exemple, considérons le cas de la régression simple, modèle (1.1). Voyons deux cas particuliers significatifs :

- si $Z_i = i$ pour $i = 1, \cdots, n$, alors $X' \cdot X = \begin{pmatrix} n & n(n+1)/2 \\ n(n+1)/2 & n(n+1)(2n+1)/6 \end{pmatrix}$, donc on a $(X' \cdot X)^{-1} \underset{n\to+\infty}{\longrightarrow} 0$ (Hypothèse (A3)), et ainsi $\widehat{\theta}^n$ converge vers θ. Cependant, on n'a pas l'existence d'un $\alpha > 0$ tel que $\frac{1}{n^\alpha}(X' \cdot X) \underset{n\to+\infty}{\longrightarrow} Q$ (Hypothèse (A2)). En fait, la convergence de $\widehat{\theta}^n = (\widehat{\theta}^n_1, \widehat{\theta}^n_2)'$ vers θ est plus rapide que la vitesse classique en $\sqrt{n}$ en ce qui concerne θ_2 (coefficient de la pente de la droite de régression). Pour vérifier cela, on calcule :

  $$H^n = \frac{2}{n(n^2-1)}\left((n+1)(2n+1) + 6ij - 3(n+1)(i+j)\right)_{1\le i,j\le n} \implies \|H^n\| = \frac{2(2n-1)}{n(n+1)},$$

  donc $\|H^n\| \underset{n\to+\infty}{\longrightarrow} 0$. La condition (A1) est donc vérifiée, et on déduit du théorème 8.2 que la vitesse de convergence est en $\sqrt{n}$ pour $\widehat{\theta}^n_1$ et en n pour $\widehat{\theta}^n_2$.

- si $Z_i = 1/i$ pour $i = 1, \cdots, n$, alors $X' \cdot X = \begin{pmatrix} n & \sum 1/i \\ \sum 1/i & \sum 1/i^2 \end{pmatrix}$. En utilisant le fait que $\sum 1/i^2$ converge (vers $\pi^2/6$), que $\sum 1/i \sim \log n$ pour $n \to \infty$, on montre que $(X' \cdot X)^{-1}$ ne converge pas vers 0. De plus, on montre que

  $$\|H^n\| = \frac{n + \sum 1/i^2 - 2\sum 1/i}{n\sum 1/i^2 - \left(\sum 1/i\right)^2},$$

  donc, $\|H^n\| \underset{n\to+\infty}{\longrightarrow} \frac{6}{\pi^2}$. Dans ce cas, il n'y a donc pas convergence de $\widehat{\theta}^n$ vers θ. En revanche, si on écrit que $\theta = (\theta_1, \theta_2)'$, on montre qu'il y a convergence de $\widehat{\theta}^n_1$ vers θ_1 (ordonnée à l'origine de la droite de régression), mais pas de $\widehat{\theta}^n_2$ vers θ_2.

3 Exercices

Exercice 8.1

(*) Soit le modèle d'analyse de la variance à un facteur équirépété. On suppose que le modèle est non gaussien mais vérifie l'hypothèse **H**. On suppose de plus que le nombre I de niveaux du facteur est constant, mais que le nombre de répétitions K tend vers l'infini. Montrer directement par une utilisation du Théorème de la Limite Centrale la normalité asymptotique de $\widehat{\theta}^n$ et la convergence de $\widehat{\sigma}^{2,n}$ vers σ^2.

Exercice 8.2

(**) Soit le modèle de régression linéaire simple de Y par Z, muni des postulats **P1-4** de paramètres μ, β et σ^2. Cependant, croyant plutôt que le modèle est quadratique, on utilise les estimateurs $\widehat{\theta}$ et $\widehat{\gamma}^{2,n}$ du modèle de régression polynômiale de degré 2 (de paramètres $\theta = (\theta_1, \theta_2, \theta_3)'$ et γ^2). On dira alors que l'on a sur-ajusté le modèle (voir chapitre 9). A-t-on $\mathbb{E}(\widehat{\theta_3}^n) = 0$; $\widehat{\theta_3}^n \xrightarrow[n\to+\infty]{\mathcal{P}} 0$; $\widehat{\gamma}^{2,n} \xrightarrow[n\to+\infty]{\mathcal{P}} \sigma^2$ quand le nombre de données n tend vers l'infini ?
Si z est une valeur donnée de la variable Z, quelle est la variance de $\widehat{Y}$ en z calculée à l'aide de l'estimateur $\widehat{\theta}$? Comparer au cas de la régression simple.

Exercice 8.3

(**) A l'inverse de l'exercice précédent, on suppose maintenant que l'on a un modèle de régression polynômiale de degré 2 muni des postulats **P1-4**, mais que pour l'analyser on utilise un modèle de régression simple. On dira alors que l'on a sous-ajusté le modèle (voir chapitre 9). En se plaçant dans le cas particulier où $Z_i = i$ pour $i = 1, \ldots, n$, vérifier que l'estimateur de la variance ne converge pas et que $|\widehat{Y}_{n+1} - Y_{n+1}| \xrightarrow[n\to+\infty]{\mathcal{P}} +\infty$.

Chapitre 9

Critères de sélection de modèles prédictifs

Ce chapitre présente la construction de différents critères, basés essentiellement sur un contraste pénalisé, permettant, après minimisation, de sélectionner un "meilleur" modèle prédictif. On étudie ensuite, lorsque le nombre de données devient grand, si le modèle retenu est bien "le vrai" modèle.

1 Introduction

Dans de nombreux problèmes concrets se pose la question du choix du modèle. En effet, au cours d'une expérience quelconque, on dispose en général de données issues de variables potentiellement explicatives, dont on ne sait pas toujours a priori si elles ont une influence réelle sur la ou les variables d'intérêt. On peut rajouter ainsi toutes les variables intervenant plus ou moins lors de l'expérience et "bricoler" un modèle de plus en plus précis pour s'adapter aux données obtenues. Un exemple simple de ce type de "bricolage" se rencontre au cours d'une régression polynômiale dans laquelle on fait croître le degré du polynôme jusqu'à obtenir une adéquation que l'on juge satisfaisante.

Un premier moyen de restreindre le nombre de variables explicatives est d'avoir une connaissance a priori, typiquement une information de nature causale, par exemple une loi physico-chimique. Un second moyen de limiter la taille du modèle est d'essayer d'obtenir un modèle explicatif, donc de ne retenir que les variables réellement explicatives. Nous avons vu précédemment différentes techniques permettant de sélectionner de telles variables, notamment par des suites de tests de Fisher ou de Student. Ce-

pendant, lorsque l'on souhaite seulement disposer d'un modèle permettant de bonnes prédictions, il n'est pas toujours performant de considérer un modèle explicatif. En effet, en toute logique, le modèle explicatif, s'il était explicitement connu (c'est-à-dire avec l'ensemble de ses paramètres), et s'il existait réellement, fournirait les meilleures prédictions possibles. Mais le fait que l'on doive se contenter d'estimer les différents paramètres du modèle engendre une incertitude sur les prédictions. Dans le cadre du modèle linéaire, cette incertitude croît avec le nombre de variables utilisées dans le modèle. Ainsi peut-on en arriver à la situation dans laquelle un "faux" modèle contenant peu de variables explicatives fournira de meilleures prédictions que le vrai modèle qui contient plus de variables. Et puis, il se peut qu'il n'existe pas vraiment de "vrai" modèle et que l'on doive se contenter de meilleurs modèles prédictifs par rapport aux données connues utilisant les variables explicatives. Ainsi, d'une manière assez générale, un critère de parcimonie, privilégiant les modèles de petite taille, présidera en partie au choix des variables dans le cadre d'un modèle prédictif.

Nous allons maintenant présenter différentes méthodes permettant une sélection d'un modèle prédictif. L'idée générale est de minimiser une distance moyenne (comme par exemple le risque quadratique) entre le vrai modèle et le modèle d'analyse. Suivant la distance considérée, suivant les approximations faites, on définira plusieurs critères, qui, une fois minimisés, fourniront le modèle estimé et donc aussi les variables sélectionnées. L'écriture de ces différents critères peut grossièrement se ramener à celle du fameux compromis biais-variance, le terme de biais correspondant à l'erreur que l'on peut commettre en oubliant certaines variables du vrai modèle, le terme de variance mesurant en quelque sorte l'incertitude liée à l'estimation des différents paramètres. Ce problème de la sélection d'un modèle prédictif se retrouve dans de nombreux autres contextes, comme par exemple, celui des réseaux de neurones.

2 Un exemple pédagogique

Pour bien comprendre la problématique ainsi que les notations, définitions, ..., présentées dans la partie théorique qui suivra, nous allons commencer par étudier un exemple "d'école" obtenu par simulation.

Considérons les sept variables suivantes qui pourront être des variables explicatives :

Variables	$Z^{(1)}$	$Z^{(2)}$	$Z^{(3)}$	$Z^{(4)}$	$Z^{(5)}$	$Z^{(6)}$	$Z^{(7)}$
Réalisations	$Z_i^{(1)}$	$Z_i^{(2)}$	$Z_i^{(3)}$	$Z_i^{(4)}$	$Z_i^{(5)}$	$Z_i^{(6)}$	$Z_i^{(7)}$
Valeurs en fonction de i	i	i^2	i^3	i^4	$\sqrt{i}$	$\frac{1}{i}$	$\log(i)$

Supposons que nous disposons des valeurs prises par 100 réalisations de ces sept variables, c'est-à-dire que l'on considère $i = 1, \cdots, 100$. Ces réalisations de variables qui seront dites **potentiellement explicatives** seront donc connues. Maintenant, nous allons simuler les réalisations Y_i de la variable à expliquer Y, en ayant **décidé arbitrairement que Y ne dépendrait linéairement que de $Z^{(2)}$, $Z^{(3)}$ et $Z^{(7)}$**, c'est-à-dire que l'on simule les réalisations :

$$Y_i = \beta_0^* + \beta_2^* \cdot Z^{(2)} + \beta_3^* \cdot Z^{(3)} + \beta_7^* \cdot Z^{(7)} + \varepsilon_i^* \quad \text{pour} \quad i = 1, \cdots, 100, \tag{9.1}$$

où $\varepsilon_i \overset{\mathcal{L}}{\sim} \mathcal{N}(0, \sigma_*^2)$ pour tout $i = 1, \cdots, 100$, les ε_i^* étant indépendants les uns les autres. Les différentes valeurs choisies des coefficients sont :

$$\beta_0^* = 5, \quad \beta_2^* = -0.03, \quad \beta_3^* = 0.0002, \quad \beta_7^* = -3 \quad \text{et} \quad \sigma_*^2 = \Big(20\Big)^2.$$

Ceci constituera ce que nous appellerons dorénavant *le vrai modèle* (nous réserverons au vrai modèle la signalisation * pour le distinguer).

Nous avons une fois pour toute simulé une réalisation de ce modèle. Le modèle 9.1 est un modèle de régression. C'est une situation classique. On supposera maintenant que sont connues les différentes valeurs Y_i et $Z_i^{(1)}, \cdots, Z_i^{(7)}$ pour $i = 1, \cdots, 100$. Les valeurs des ε_i, la variance σ_*^2, les valeurs des différents coefficients β_j^*, sont inconnus et surtout on ne sait pas quels sont parmi les coefficients β_j^* ceux qui sont nuls, c'est-à-dire que l'on ne connaît pas le "vrai" modèle (*i.e.* le choix dans $\{1, 2, 3, 4, 5, 6, 7\}$ des indices j des variables $Z^{(j)}$ intervenant vraiment).

Voici maintenant les commandes en R permettant une telle simulation et traçant également dans une figure de taille déterminée à l'avance (commandes `xlim` et `ylim`) le nuage de points et la "vraie" fonction pour $i = 1$ à $i = 150$ (que l'on appelle également prédiction optimale par moindres carrés) :

```
i=1:100;
Z1=i;Z2=i^2;Z3=i^3;Z4=i^4;Z5=sqrt(i);Z6=1/i;Z7=log(i);
epsilon=20*rnorm(100,0);
Y=5-0.03*Z2+0.0002*Z3-3*Z7+epsilon;
```

```
j=1:150;
YY=5-0.03*(j^2)+0.0002*(j^3)-3*log(j)
plot(j,YY,"l",xlim=c(-10,160),ylim=c(-150,50))
points(i,Y)
```

(la figure 2 sera tracée un peu plus bas). Pour déterminer le vrai modèle, nous avons vu au chapitre 4 une méthodologie possible à l'aide de tests de Fisher emboîtés, ce que nous avons appelé une régression pas-à-pas descendante. Si on applique cette méthode aux données précédentes, on pourra écrire les commandes suivantes :

```
library(car)
y.lm7=lm(Y~Z1+Z2+Z3+Z4+Z5+Z6+Z7)
new=data.frame(Z1=j,Z2=j^2,Z3=j^3,Z4=j^4,Z5=sqrt(j),Z6=1/j,Z7=log(j))
y.pred7=predict(y.lm7,new)
points(j,y.pred7,"l",lty="dashed")
Anova(y.lm7)
Anova(lm(Y~Z1+Z2+Z3+Z4+Z5+Z7))
Anova(lm(Y~Z1+Z2+Z3+Z4+Z5))
Anova(lm(Y~Z1+Z2+Z3+Z5))
Anova(lm(Y~Z1+Z2+Z3))
Anova(lm(Y~Z2+Z3))
y.lm2=lm(Y~Z2+Z3)
y.pred2=predict(y.lm2,new)
points(j,y.pred2,pch=20)
```

(à partir des relevés donnés par la commande `Anova`, on réalise à la main l'équivalent d'une régression descendante pas-à-pas : on enlève à chaque étape la variable possédant la P-value la plus grande et on passe à l'étape suivante jusqu'à obtenir des P-values toutes inférieures à 5%). Les lignes de commande $3-4$ et les 2 dernières permettent d'ajouter au nuage de points d'une part la fonction prédite (pour $i=1$ à $i=150$) avec le modèle contenant toutes les variables, et d'autre part, la fonction prédite avec le modèle sélectionné par régression pas-à-pas descendante.

Pour "notre" simulation, nous avons obtenu (extraits des résultats) :

```
>Anova(y.lm7)
Anova Table (Type II tests)

Response: Y
          Sum Sq Df F value Pr(>F)
```

```
Z1             8  1  0.0220 0.8824
Z2            55  1  0.1496 0.6999
Z3           100  1  0.2719 0.6033
Z4           117  1  0.3189 0.5736
Z5             1  1  0.0024 0.9608
Z6             1  1  0.0024 0.9608
Z7        0.2692  1  0.0007 0.9784
Residuals  33674 92

    :           :           :           :
    :           :           :           :

> Anova(lm(Y~Z2+Z3))
Anova Table (Type II tests)

Response: Y
          Sum Sq Df F value    Pr(>F)
Z2         33797  1  93.941 6.661e-16 ***
Z3         14938  1  41.522 4.509e-09 ***
Residuals  34897 97
```

On comprend donc qu'après avoir utilisé un modèle contenant toutes les variables potentiellement explicatives, on recommence avec un modèle ne contenant plus la variable $Z^{(7)}$ car sa P-value ($\simeq 0.978$) est la plus élevée. Finalement, on aboutit à un modèle contenant uniquement les variables $Z^{(2)}$ et $Z^{(3)}$. (toujours dans le cas de cette simulation...). La régression pas-à-pas descendante a donc "oublié" la variable $Z^{(7)}$ par rapport au vrai modèle (9.1).

Nous allons maintenant utiliser des *critères appelés Cp de Mallows, AIC et BIC* qui permettent de déterminer un meilleur modèle prédictif (dans un sens précis qui sera examiné par la suite). Pour l'instant, nous ne nous intéressons pas à l'expression exacte de ces critères, qui peuvent être calculés pour chaque modèle choisi : ainsi, si on choisit par exemple les variables $Z^{(1)}$ et $Z^{(3)}$, on peut calculer les valeurs de ces 3 critères pour le modèle linéaire de Y par rapport à ces deux variables. L'idée sera de minimiser ces critères sur un ensemble de ces modèles pour sélectionner un "meilleur" modèle prédictif (en un certain sens). Voyons cela plus précisément à partir des commandes R appliquées sur le jeu de données, en commençant par le critère Cp :

```
Z=matrix(c(Z1,Z2,Z3,Z4,Z5,Z6,Z7),ncol=7);
colnames(Z)=c("Z1","Z2","Z3","Z4","Z5","Z6","Z7");
r=leaps(Z,Y);
```

```
r$whi;
r$Cp;
t=(r$Cp==min(r$Cp));
colnames(Z)[r$whi[t]]
y.pred.cp=predict(lm(Y~Z2+Z4),new)
points(j,y.pred.cp,pch=19)
```

Voici un extrait des résultats obtenus :

```
>r$whi
      1     2     3     4     5     6     7
1  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
1 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
1 FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
1 FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
1 FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
2 FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
:   :      :     :    :     :     :     :

> r$Cp;
 [1]   9.219291  28.698723  40.154159  91.677566 104.267979 138.767126 343.985273
 [8]   1.115052   1.341950   5.961345   9.396759  10.222900  10.791692  10.973616
[15]  10.990467  11.099619  11.297052   2.658281   2.811358   3.009853   3.048201
[22]   3.053448   3.107051   3.113925   3.206643   3.253720   3.869152   2.945192
[29]   3.082472   3.367223   3.390150   3.415472   3.503617   3.561068   3.611091
  :       :          :          :          :          :          :          :

> colnames(Z)[r$whi[t]]
[1] "Z2" "Z4"
```

Pour chaque modèle (donné ici par la présence (`TRUE`) ou non (`FALSE`) de la variable potentiellement explicative), on calcule la valeur du critère Cp (par exemple, pour la première ligne, le modèle contient comme seule variable explicative $Z^{(1)}$, et la valeur du Cp associée est $\simeq 9.219291$).

Les deux dernières lignes de commande permettent d'obtenir explicitement le modèle sélectionné en minimisant le critère Cp (dans le cas présent, on aura obtenu un modèle

différent du vrai modèle et du modèle sélectionné par régression pas-à-pas descendante, puisque les variables explicatives retenues sont $Z^{(2)}$ et $Z^{(4)}$). Voyons maintenant l'obtention des modèles sélectionnés par les deux autres critères :

```
ZZ=as.data.frame(Z);
y.lm=lm(Y~.,data=ZZ);
y.bic=stepAIC(y.lm,k=log(100))
y.aic=stepAIC(y.lm,k=2,trace=FALSE)
```

Pour calculer le critère BIC, on utilise la commande `stepAIC`, avec un paramètre $k = \log(n)$, alors que par défaut, $k = 2$, ce qui correspond à la valeur de k pour le critère AIC (attention ! dans la suite, les résultats s'écriront avec AIC, alors que, si par exemple $k = \log(n)$, ce pourra être BIC). Notez que le rajout de l'option `trace=FALSE` permet d'avoir le modèle final sélectionné sans les étapes intermédiaires. Voici le résultat de ces commandes :

```
> y.bic=stepAIC(y.lm,k=log(100))
Start:  AIC= 618.77
 Y ~ Z1 + Z2 + Z3 + Z4 + Z5 + Z6 + Z7

       Df Sum of Sq   RSS   AIC
- Z7    1    0.2692 33674   614
- Z6    1         1 33675   614
- Z5    1         1 33675   614
- Z1    1         8 33682   614
- Z2    1        55 33729   614
- Z3    1       100 33774   614
- Z4    1       117 33791   615
<none>              33674   619

:   :      :     :    :     :
:   :      :     :    :     :

Step:  AIC= 599.31
 Y ~ Z2 + Z3

       Df Sum of Sq   RSS   AIC
<none>              34897   599
- Z3    1     14938 49835   630
- Z2    1     33797 68694   662
```

```
> y.aic

Coefficients:
(Intercept)           Z1           Z2           Z3           Z4           Z5
  6.405e+01    1.452e+01   -2.872e-01    2.856e-03   -1.014e-05   -5.962e+01
```

Ainsi, avec ce jeu de données, les critères BIC et AIC auront sélectionné respectivement les modèles `Y` $\sim$ `Z2 + Z3` (qui est aussi le modèle obtenu par par régression pas-à-pas descendante) et `Y` $\sim$ `Z1 + Z2 + Z3 + Z4 + Z5`, qui sont différents du modèle choisi à partir du critère Cp de Mallows (ceci n'est pas une règle générale, car avec le même exemple mais d'autres simulations du bruit, on obtient assez souvent le même modèle sélectionné par les 3 critères, qui est également le même que celui fourni par régression pas-à-pas descendante...). Voici une synthèse des résultats obtenus :

Critère	Modèle sélectionné
Modèle complet	`Y` $\sim$ `Z1 + Z2 + Z3 + Z4 + Z5 + Z6 + Z7`
Vrai modèle	`Y` $\sim$ `Z2 + Z3 + Z7`
Régression pas-à-pas descendante	`Y` $\sim$ `Z2 + Z3`
Cp de Mallows	`Y` $\sim$ `Z2 + Z4`
BIC	`Y` $\sim$ `Z2 + Z3`
AIC	`Y` $\sim$ `Z1 + Z2 + Z3 + Z4 + Z5`

Il ne nous reste plus qu'à comparer les prédictions suivant les modèles sélectionnés. Cela peut se faire à partir du graphe 2 :

De cette figure, on s'aperçoit que le modèle contenant toutes les variables $Z^{(j)}$ sans discrimination conduit à des prédictions catastrophiques (ce phénomène est appelé "sur-ajustement"), alors que le modèle sélectionné par régression pas-à-pas descendante ou par le critère BIC, amène à des prédictions clairement plus correctes (de même, dans une moindre mesure pour la prédiction par le critère Cp, la sélection par le critère AIC conduisant, sur ce jeu de données, à d'assez médiocres prédictions). Ceci n'apparaissait pas si l'on se contentait d'obtenir la meilleure adéquation possible (qui est toujours obtenu par le modèle contenant toutes les variables). **Il nous faut donc retenir de tout ceci :**

- **en vue de prédictions, il faut toujours essayer d'éliminer le plus de variables potentiellement explicatives en choisissant des modèles d'analyse les plus parcimonieux ;**

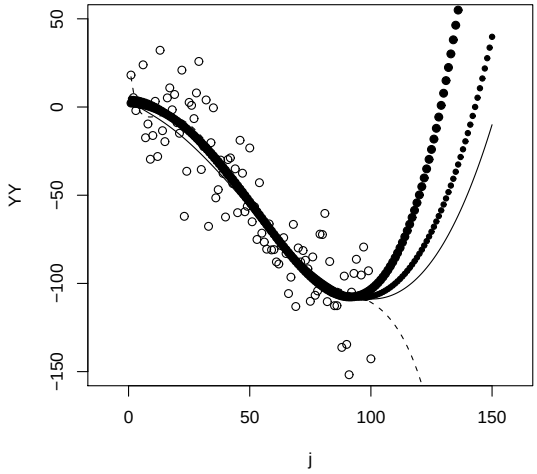


FIGURE 9.1 – Nuage de points et 1/ la vraie fonction (en ligne continue), 2/ la fonction estimée avec le modèle contenant toutes les variables (en ligne pointillé), 3/ la fonction estimée à partir du critère Cp de Mallows (en gros cercles foncés) et 4/ la fonction estimée à partir du modèle sélectionné par régression pas-à-pas descendante et par le critère BIC (en petits cercles foncés)

- **la recherche d'un "bon" modèle d'analyse peut se faire avec différentes méthodes qui ne conduisent pas forcément à retrouver le vrai modèle.**

3 Présentation générale et définition

Nous allons maintenant généraliser l'exemple précédent au cadre d'une régression linéaire quelconque.

Hypothèses et notations : On suppose que l'on dispose des n résultats d'une expérience concernant une variable d'intérêt quantitative réelle Y, variable dite à expliquer, et que k (nombre fixe ne dépendant pas de n) variables quantitatives réelles $Z^{(i)}$, $i = 1, \cdots, k$ peuvent être utilisées pour expliquer Y. Pour chaque expérience $j = 1, \cdots, n$, on connaît Y_j et $Z_j^{(1)}, \cdots, Z_j^{(k)}$. Par la suite, on désignera également par Y le vecteur colonne de taille n, $Y = (Y_1, \cdots, Y_n)'$, ainsi que par $Z^{(i)}$ le vecteur colonne $(Z_1^{(i)}, \cdots, Z_n^{(i)})'$, pour $i = 1, \cdots, k$. Enfin, on définit $\mathcal{M}$ un ensemble de modèles, qui est une famille de sous-ensembles de $\{1, \cdots, k\}$. Deux exemples sont à retenir :

- $\mathcal{M} = \mathcal{P}(\{1, \cdots, k\})$: famille exhaustive de modèles ;
- $\mathcal{M} = \{1, \cdots, j\}_{1 \leq j \leq k}$: famille hiérarchique de modèles (par exemple, des modèles polynômiaux) appelée parfois également famille de modèles emboîtés.

Par la suite, pour $m \in \mathcal{M}$, on notera $|m| = \text{Card}(m)$ et $X_{(m)}$. On dispose donc de $|m|$ variables potentiellement explicatives. Pour ne pas mettre sur le même plan la constante (estimée par l'intercept) et ces variables, on considérera la matrice de taille $(n, |m| + 1)$ s'écrivant sous la forme :

$$X_{(m)} = \begin{pmatrix} 1 & Z_1^{(i_1)} & Z_1^{(i_2)} & \cdots & Z_1^{(i_{|m|})} \\ 1 & Z_2^{(i_1)} & Z_2^{(i_2)} & \cdots & Z_2^{(i_{|m|})} \\ \vdots & \vdots & & \vdots & \\ 1 & Z_n^{(i_1)} & Z_n^{(i_n)} & \cdots & Z_n^{(i_{|m|})} \end{pmatrix} \quad \text{lorsque} \quad m = \{i_1, \cdots, i_{|m|}\}.$$

On supposera également par la suite que pour tout $m \in \mathcal{M}$ la matrice $X_{(m)}$ est régulière (donc son rang est $|m| + 1$). Notez que le $+1$ vient de la constante qui est supposée être présente systématiquement dans tous les modèles.

Hypothèses sur le vrai modèle : On suppose qu'il existe $m^* \in \mathcal{M}$, inconnu, tel que le vrai modèle s'écrive :

$$Y = \mu^* + \varepsilon^* = X_{(m^*)} \cdot \theta_{(m^*)} + \varepsilon^*, \quad \text{où} \quad \varepsilon^* \sim \mathcal{N}(0, \sigma_*^2 \cdot I_n), \tag{9.2}$$

avec $\mu^* \in \mathbb{R}^n$, I_n étant la matrice identité de taille n, $\theta_{(m^*)}$ un vecteur colonne réel de taille $|m^*|+1$ et ayant **toutes ses coordonnées non nulles**, ε^* et σ_* des paramètres inconnus.

Remarque : Observons que les ε_i^* sont indépendants et de même variance.

Modèles d'analyse : Pour modéliser l'expérience et essayer d'identifier le vrai modèle, on utilise la famille de modèles suivante, qui est en correspondance avec $\mathcal{M}$, soit

$$Y = \mu + \varepsilon = X_{(m)} \cdot \theta_{(m)} + \varepsilon, \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \cdot I_n), \tag{9.3}$$

avec $m \in \mathcal{M}$, $\mu \in \mathbb{R}^n$, $\theta_{(m)} \in \mathbb{R}^{|m|+1}$. Par la suite, on appellera modèle m le modèle d'analyse dans lequel on choisit les variables correspondant à m.

Pour préciser la modélisation, nous utiliserons les définitions suivantes :

Définition : On suppose que le modèle d'analyse est $m \in \mathcal{M}$. Alors :

- si $m = m_k = \{1, \cdots, k\}$, on dit que le modèle est complet, c'est-à-dire que toutes les variables explicatives disponibles sont considérées.
- si $m^* \subset m$ avec $m \neq m^*$, on dit que le modèle est sur-ajusté.

- si $m \subset m^*$ avec $m \neq m^*$, on dit que le modèle est faux ;

Rappelons que chaque modèle correspond à un choix parmi l'ensemble des variables explicatives, et qu'il y a donc potentiellemnt des variables explicatives superflues. En cas de sur-ajustement, c'est-à-dire s'il y a des variables superflues, un modèle sur-ajusté est un modèle contenant toutes les variables du vrai modèle, plus un certain nombre de variables superflues. Un faux-modèle est donc typiquement un modèle où les variables du vrai modèle n'ont pas toutes été choisies et où certaines variables superflues ont pu être choisies. Un cas particulier est celui du sous-ajustement correspondant à un faux modèle ne contenant aucune variable superflue.
Exemple : En reprenant l'exemple (dit "d'école") précédent de vrai modèle (9.1), la famille de modèle exhaustive est $\mathcal{M} = \{A\,;\, A \subset \{1,2,3,4,5,6,7\}\}$, une famille hiérarchique pourra être $\mathcal{M}_H = \{\{2\},\{2,3\},\{2,3,1\},\{2,3,1,5\},\{1,2,3,4,5\},\{1,2,3,4,5,6\},\{1,2,3,4,5,6,7\}\}$. Le modèle d'analyse $Y = \beta_0 + \beta_2 \cdot Z^{(2)} + \varepsilon$ est un faux modèle sous-ajusté, quand le modèle d'analyse $Y = \beta_0 + \beta_1 \cdot Z^{(1)} + \beta_2 \cdot Z^{(2)} + \beta_3 \cdot Z^{(3)} + \beta_5 \cdot Z^{(5)} + \beta_7 \cdot Z^{(7)}\varepsilon$ est lui sur-ajusté. Enfin, le modèle $Y = \beta_0 + \beta_2 \cdot Z^{(5)} + \varepsilon$ est un faux modèle.

Rappels : On suppose que l'on choisit le modèle d'analyse $m \in \mathcal{M}$. Alors :

i. L'estimateur de $\theta_{(m)}$ par moindres carrés ordinaires est :

$$\widehat{\theta}_{(m)} = ((X_{(m)})' \cdot X_{(m)})^{-1} \cdot (X_{(m)})' \cdot Y \quad \text{d'où} \quad \widehat{Y}_{(m)} = X_{(m)} \cdot \widehat{\theta}_{(m)}.$$

ii. Un estimateur de σ^2 (non biaisé lorsque $m^* \subset m$) est :

$$\widehat{\sigma}^2_{(m)} = \frac{1}{(n - |m|)} ||Y - \widehat{Y}_{(m)}||^2 = \frac{1}{(n - |m|)} (Y - \widehat{Y}_{(m)})' \cdot (Y - \widehat{Y}_{(m)}).$$

iii. Pour des raisons de simplifications de calculs, nous utiliserons également l'estimateur biaisé $\tilde{\sigma}^2_{(m)}$ de σ^2 (que l'on obtient par maximum de vraisemblance dans le cas gaussien lorsque le modèle d'analyse est le vrai modèle), et qui a pour expression :

$$\tilde{\sigma}^2_{(m)} = \frac{1}{n} ||Y - \widehat{Y}_{(m)}||^2.$$

4 Distances entre deux modèles

Pour mesurer l'écart entre le vrai modèle (avec m^*) et un modèle d'analyse (soit $m \in \mathcal{M}$), nous utiliserons le critère du risque quadratique et celui de la dissemblance de Kullback.

4.1 Risque quadratique

Rappelons ici la définition du risque quadratique, que l'on rencontre usuellement pour, par exemple, montrer la convergence d'un estimateur paramétrique. Nous l'utiliserons ici de façon assez naturelle puisque finalement mesurer une distance entre m et m^* c'est aussi mesurer une distance entre $\mu^* = X_{(m^*)} \cdot \theta_{(m^*)}$ et son estimateur $\widehat{Y}_{(m)}$ par le modèle d'analyse m.

Définition 9.1 *Le risque quadratique entre les modèles m et m^* est :*

$$R(m, m^*) = \mathbb{E}\left(||\mu^* - \widehat{Y}_{(m)}||^2\right) = \mathbb{E}\left(||X_{(m^*)} \cdot \theta_{(m^*)} - X_{(m)} \cdot \widehat{\theta}_{(m)}||^2\right).$$

Remarquons que dans le cas où $m = m^*$, comme $\mathbb{E}\left(\widehat{Y}_{(m)}\right) = \mu^*$, on a plus précisément

$$R(m, m^*) = \mathbb{E}\left(||\widehat{Y}_{(m)}||^2\right) - ||\mu^*||^2.$$

4.2 Dissemblance de Kullback

A chaque modèle d'analyse, qui, en général, est un faux modèle, on peut faire correspondre la mesure de probabilité de Y en procédant comme si ce modèle d'analyse était réellement le vrai modèle. On fera donc correspondre la loi de $X_{(m)} \cdot \widehat{\theta}_{(m)} + \widehat{\varepsilon}$. On pourra ainsi mesurer l'écart existant entre la loi du vrai modèle (loi paramétrée par des paramètres inconnus) et la loi engendrée par le modèle d'analyse. Pour mesurer cet écart, un outil souvent utilisé est la dissemblance de Kullback :

Définition 9.2 *Soit $\mathbb{P}$ et $\mathbb{P}^*$ deux mesures de probabilité dominées par une même mesure ν (ce qui est toujours le cas) . La dissemblance de Kullback de $\mathbb{P}^*$ sur $\mathbb{P}$ est :*

$$K(\mathbb{P}^*, \mathbb{P}) = \mathbb{E}_{\mathbb{P}^*}\left(\log \frac{d\mathbb{P}^*}{d\mathbb{P}}\right).$$

Si $f = \dfrac{d\mathbb{P}}{d\nu}$ et $f^* = \dfrac{d\mathbb{P}^*}{d\nu}$, alors, $K(\mathbb{P}^*, \mathbb{P}) = \begin{cases} \displaystyle\int f^* \log \frac{f^*}{f} d\nu & \text{si } \mathbb{P}^* << \mathbb{P}\,; \\ +\infty & \text{sinon.} \end{cases}$

Remarquons en premier lieu la non symétrie de $K(.,.)$. C'est pour cette raison que l'on préférera parler de dissemblance plutôt que de distance. Cependant, cette dissemblance vérifie comme pour une distance "classique" la propriété suivante :

$K(\mathbb{P}, \mathbb{P}^*) \geq 0$ pour toutes mesures $\mathbb{P}$ et $\mathbb{P}^*$, et $K(\mathbb{P}, \mathbb{P}^*) = 0$ si et seulement si $\mathbb{P} = \mathbb{P}^*$ (ceci peut être démontré par des arguments de convexité).

Dans **le cadre gaussien** initialement proposé, on considère des densités par rapport à la mesure de Lebesgue et, **si on suppose que l'on connaît les paramètres μ et σ du modèle d'analyse** (hypothèse qui n'a pas de réalité, mais qui va nous permettre une première étude de la dissemblance), on a pour $y \in \mathbb{R}$:

$$\begin{aligned} f^*(y) &= \frac{1}{(2\pi\sigma_*^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_*^2}(y-\mu^*)'.(y-\mu^*)\right) \\ f(y) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)'.(y-\mu)\right) \end{aligned}$$

Ainsi, en considérant la dissemblance de $\mathbb{P}^*$ sur $\mathbb{P}$, ce que l'on notera également $K(f^*, f)$, on obtient :

$$K(f^*, f) = \frac{2}{n}\left(\log\left(\frac{\sigma_*^2}{\sigma^2}\right) - \frac{1}{n}\mathbb{E}\left[\frac{1}{\sigma_*^2}||Y-\mu^*||^2 - \frac{1}{\sigma^2}||Y-\mu||^2\right]\right).$$

Mais comme $\mathbb{E}(Y) = \mu^*$, alors $\mathbb{E}\left[(Y-\mu^*)' \cdot (Y-\mu^*)\right] = \mathbb{E}\left(||\varepsilon^*||^2\right) = n\sigma_*^2$.

De plus, $||Y-\mu||^2 = ||Y-\mu^*||^2 + 2(Y-\mu^*)' \cdot (\mu^*-\mu) + ||\mu^*-\mu||^2$, donc avec ce qui précède,

$$K(f^*, f) = \frac{n}{2}\left(\log\left(\frac{\sigma^2}{\sigma_*^2}\right) + \frac{\sigma_*^2}{\sigma^2} - 1\right) + \frac{1}{2\sigma^2}||\mu^*-\mu||^2.$$

Par la suite, pour mesurer la véritable dissemblance entre m et m^*, on remplacera μ et σ^2 par leurs estimations respectives, $\widehat{\mu}_{(m)}$ et $\tilde{\sigma}^2_{(m)}$, qui dépendent de Y.

5 Cinq critères pour sélectionner un modèle

Les deux mesures entre modèles présentées ci-dessus vont naturellement offrir des moyens de définir des critères de choix de modèles par leur minimisation.

5.1 Le critère CP de Mallows comme minimisation du risque quadratique

Une idée naturelle pour sélectionner le "meilleur modèle" est de directement minimiser le risque quadratique. Dans le cadre dans lequel nous nous sommes placés, pour un modèle $m \in \mathcal{M}$, le risque quadratique s'écrit :

$$\begin{aligned} R(m, m^*) &= \mathbb{E}\left[||\widehat{Y}_{(m)} - \mu^*||^2\right] = \mathbb{E}\left[||X_{(m)}.\widehat{\theta}_{(m)} - \mu^*||^2\right] \\ &= \mathbb{E}\left[||X_{(m)}.\widehat{\theta}_{(m)} - \mu^*_{(m)}||^2\right] + \mathbb{E}\left[||\mu^* - \mu^*_{(m)}||^2\right], \end{aligned}$$

en appelant $\mu^*_{(m)}$ (respectivement $\varepsilon^*_{(m)}$) le projeté orthogonal de μ^* (respectivement ε^*) sur le sous-espace vectoriel $[X_{(m)}]$ (espace de dimension $(|m|+1)$), la dernière égalité découlant du Théorème de Pythagore. Par linéarité de la projection et de l'espérance, il est clair que $\mathbb{E}\left[\widehat{Y}_{(m)}\right] = \mu^*_{(m)}$. En conséquence, comme $\widehat{Y}_{(m)}$ est la projection orthogonale de Y sur $[X_{(m)}]$, alors, en notant $P_{(m)}$ la matrice de projection orthogonale dans $\mathbb{R}^n$ sur $[X_{(m)}]$ et en utilisant $\mathrm{Tr}(A \cdot B) = \mathrm{Tr}(B \cdot A)$) :

$$\begin{aligned} \mathbb{E}\left[||\widehat{Y}_{(m)} - \mu^*||^2\right] &= \mathbb{E}\left[||\varepsilon^*_{(m)}||^2\right] + \mathbb{E}\left[||\mu^* - \mu^*_{(m)}||^2\right] \\ &= \mathbb{E}\left[\mathrm{Tr}((\varepsilon^*_{(m)})' \cdot \varepsilon^*_{(m)})\right] + ||\mu^* - \mu^*_{(m)}||^2 \\ &= \mathrm{Tr}\left(P'_{(m)} \cdot P_{(m)} \cdot \mathbb{E}\left[\varepsilon^* \cdot (\varepsilon^*)'\right]\right) + ||\mu^* - \mu^*_{(m)}||^2 \\ &= \sigma^2_* \mathrm{Tr}\left(P'_{(m)} \cdot P_{(m)}\right) + ||\mu^* - \mu^*_{(m)}||^2 \\ &= (|m|+1) \cdot \sigma^2_* + ||\mu^* - \mu^*_{(m)}||^2, \end{aligned}$$

le membre de droite de cette égalité ne nécessitant pas d'espérance puisqu'il est déterministe. Notons que $(|m|+1) \cdot \sigma^2_*$ aurait pu être obtenu plus rapidement dans un cadre gaussien en appliquant le Théorème de Cochran.

Remarque : On obtient ainsi l'expression générale du risque quadratique pour un modèle quelconque. Remarquons qu'il est composé de deux parties comportant des paramètres inconnus. Il n'est donc pas possible de minimiser directement ce risque. On observe également que la première partie est un terme de variance qui augmente avec la dimension du modèle choisi, quant la deuxième partie est un terme de biais, qui diminue quand augmente la dimension du modèle, jusqu'à s'annuler quand le modèle d'analyse m contient le vrai modèle m^*. On rencontre donc ici un compromis biais-variance, qui heuristiquement s'explique par le fait que plus la dimension du modèle augmente, meilleur est l'ajustement du modèle aux données, mais plus importante est la variabilité du modèle (donc la possible erreur lors d'une prédiction).

Pour minimiser le risque quadratique $R(m, m^*)$ et ainsi sélectionner un "meilleur modèle" suivant ce critère, on est amené à utiliser des estimations des parties biais et variance. En effet, on a :

$$\begin{aligned}
\mathbb{E}\left[||Y - \widehat{Y}_{(m)}||^2\right] &= \mathbb{E}\left[||Y - \mu^*_{(m)}||^2\right] - \mathbb{E}\left[||\widehat{Y}_{(m)} - \mu^*_{(m)}||^2\right] \quad \text{(Pythagore)} \\
&= \mathbb{E}\left[||Y - \mu^* + \mu^* - \mu^*_{(m)}||^2\right] - \mathbb{E}\left[||\widehat{Y}_{(m)} - \mu^*_{(m)}||^2\right] \\
&= \mathbb{E}\left[||Y - \mu^*||^2\right] + ||\mu^* - \mu^*_{(m)}||^2 - (|m|+1)\cdot\sigma^2_* \\
&= (n - (|m|+1))\cdot\sigma^2_* + ||\mu^* - \mu^*_{(m)}||^2
\end{aligned}$$

(ici encore cette expression est vraie dans un cadre non gaussien). Par suite, on va estimer la partie biais $||\mu^* - \mu^*_{(m)}||^2$ du risque quadratique, puisque celle-ci dépend du paramètre μ^* inconnu. Cela peut se faire par substitution de la manière suivante :

$$\begin{aligned}
R(m, m^*) &= (|m|+1)\cdot\sigma^2_* + \mathbb{E}\left[||Y - \widehat{Y}_{(m)}||^2\right] - (n - (|m|+1))\cdot\sigma^2_* \\
\text{soit } \frac{R(m, m^*)}{n\cdot\sigma^2_*} &= \frac{2|m|+1}{n} - 1 + \frac{1}{n\cdot\sigma^2_*}\,\mathbb{E}\left[||Y - \widehat{Y}_{(m)}||^2\right] \\
\text{et donc } \frac{R(m, m^*)}{n\sigma^2_*} &\simeq \frac{2|m|+1}{n} - 1 + \frac{\tilde{\sigma}^2_{(m)}}{\sigma^2_*}
\end{aligned}$$

en estimant $\mathbb{E}\left[||Y - \widehat{Y}_{(m)}||^2\right]$ par $||Y - \widehat{Y}_{(m)}||^2$, ce qui n'est pas illégitime car, par exemple, lorsque $m* \subset m$, alors $\tilde{\sigma}^2_{(m)}$ est un estimateur convergent de $\dfrac{\mathbb{E}\,||Y - \widehat{Y}_{(m)}||^2}{n}$ lorsque $n \to \infty$ (en supposant que k, le nombre total de régresseurs, ne dépende pas de n).

On conserve cependant un terme inconnu dans cette estimation de la distance quadratique, à savoir σ_*, issu de la partie variance du risque quadratique. Or on sait que $m^* \in \mathcal{M}$, donc en considérant $m_k = \{1, \cdots, k\}$, $\tilde{\sigma}_{(m_k)}$ est un estimateur convergent de σ_* (toujours en supposant que k ne varie pas avec n). Ainsi nous utiliserons pour minimiser notre distance quadratique une minimisation du critère appelé *Cp de Mallows* et valant pour $m \in \mathcal{M}$, :

$$Cp(m) = \frac{\tilde{\sigma}^2_{(m)}}{\tilde{\sigma}^2_{(m_k)}} + 2\,\frac{|m|}{n}$$

(on peut trouver d'autres expressions de ce critère, qui sont en fait les mêmes à une constante multiplicative ou additive près, ce qui n'importe pas pour sa minimisation). On sélectionnera donc un modèle $\widehat{m}$ tel que :

$$\widehat{m}_{CP} = \text{Argmin}_{m\in\mathcal{M}}\{Cp(m)\}.$$

Ce critère a été introduit par Mallows en 1967 (voir Mallows [42] et [43]). Nous étudierons ses propriétés, notamment asymptotiques, un peu après.

5.2 Le critère PRESS comme une minimisation de sommes de risques quadratiques

Le critère du Cp de Mallows trouve sa légitimation dans plusieurs approximations successives du risque quadratique, ceci bien-sûr parce que le vrai modèle est inconnu. Il est aussi possible de directement minimiser l'erreur quadratique que l'on fait quand, pour un modèle donné m, on estime chaque valeur obtenue Y_i par une prédiction obtenue à partir des autres données $(Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n)$. Cette démarche est assez générique en statistique, on la retrouve notamment en estimation non-paramétrique, en filtrage exponentiel,... On l'appelle **validation croisée**, "cross-validation" en anglais, et ici le principe décrit est celui où l'on prédit en délaissant une seule donnée à la fois ("leave-one-out" en anglais) mais il eut aussi été possible de faire de délaisser des groupes de données.
Plus précisément, notons Y^{-i} le vecteur $(Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n)$ pour $i = 2, \cdots, n-1$, $Y^{-1} = (Y_2, \cdots, Y_n)$ et $Y^{-n} = (Y_1, \cdots, Y_{n-1})$. De même on considère $X_{(m)}^{-i}$ la matrice $X_{(m)}$ à laquelle on a retiré la ligne i. Pour le modèle m on calcule l'estimation du vecteur θ à partir de Y^{-i} :

$$\widehat{\theta}_{(m)}^{-i} := \left((X_{(m)}^{-i})' X_{(m)}^{-i}\right)^{-1} (X_{(m)}^{-i})' Y^{-i}.$$

Le critère PRESS (Predicted REsidual Sums of Squares) pour le modèle m sera alors :

$$PRESS(m) := \sum_{i=1}^{n} \left(Y_i - (X_{(m)}\, \widehat{\theta}_{(m)}^{-i})_i\right)^2,$$

où $(X_{(m)}\, \widehat{\theta}_{(m)}^{-i})_i$ est la i-ème coordonnée du vecteur $X_{(m)}\, \widehat{\theta}_{(m)}^{-i}$ et représente la prédiction de Y_i par moindres carrés ordinaires à partir de Y^{-i}. Ainsi le critère représente PRESS pour un modèle m donné une distance quadratique entre Y (fournit par le modèle m^*) et une prédiction de Y (notons que cette prédiction est obtenue par transformation linéaire de Y). Alors on peut définir :

$$\widehat{m}_{PRESS} = \text{Argmin}_{m \in \mathcal{M}} \{PRESS(m)\}.$$

5.3 Le critère R^2 ajusté comme une minimisation de risque quadratique

Le coefficient de détermination R^2 (voir sa définition en (1.4)) mesure d'une certaine manière l'adéquation entre le modèle et les données observées puisque pour un

modèle m :

$$R^2(m) = 1 - \frac{\|Y - \widehat{Y}_{(m)}\|^2}{\|Y - \overline{Y}\|^2}.$$

Notons que $\overline{Y} = \widehat{Y}_{\mathbb{1}\!\mathrm{I}}$, projection de Y sur le sous-espace engendré par le vecteur $\mathbb{1}\!\mathrm{I}$ (correspondant à l'intercept de la régression). Dans le cadre d'un problème du problème de test où H_0 : "aucune variable n'intervient" contre H_1 : "toutes les variables du modèle m interviennent", la statistique de test de Fisher est

$$\widehat{F} = \frac{n - |m| - 1}{|m|} \frac{\|Y - \overline{Y}\|^2 - \|Y - \widehat{Y}_{(m)}\|^2}{\|Y - \widehat{Y}_{(m)}\|^2},$$

donc on obtient après quelques calculs que $R^2(m) = 1 - \dfrac{1}{1 + \frac{p}{n-p-1}\widehat{F}}$ où l'on sait que sous H_0 et dans le cadre gaussien, $\widehat{F}$ suit une loi de Fisher $F(p, n-p-1)$.
Ce critère $R^2(m)$, si on le maximise apparaît donc comme une minimisation de $\|Y - \widehat{Y}_{(m)}\|^2$ (puisque le dénomisateur ne dépend pas de m). Mais il est clair que $\|Y - \widehat{Y}_{(m)}\|^2 = \|P_{[X_{(m)}]^\perp} Y\|^2$ décroît pour une suite emboitée de modèle, donc en aucun cas le coefficient R^2 ne peut être utilisé pour comparer des modèles ayant un nombre de variables différent. Une maximisation du R^2 conduit à coup sûr à chosir le modèle complet m_k. En revanche, pour des modèles de même cardinal $|m|$ ce coefficient peut être utilisé pour choisir un modèle optimal.
Il est possible d'améliorer le coefficient R^2 pour permettre de sélectionner des modèles comportant un nombre différent de variables explicatives en définissant le coefficient R^2 ajusté, noté R^2_{Aju}. Nous allons juste présenter l'intuition qui préside à cette définition. Dans la définition de $R^2(m)$, le modèle intervient dans la somme de carrés du numérateur. Un premier moyen de limiter l'influence de la taille du modèle est de diviser cette somme par $n - |m| - 1$ qui est la dimension du sous-espace de projection. Ainsi l'espérance du numérateur est σ^2, et ne dépend plus de m. On peut par homogénéisation faire la même chose avec le dénominateur même s'il ne dépend pas de m. Aussi aboutit-on à la définition :

$$R^2_{Aju}(m) = 1 - \frac{\frac{1}{n-|m|-1}\|Y - \widehat{Y}_{(m)}\|^2}{\frac{1}{n-1}\|Y - \overline{Y}\|^2}.$$

L'écriture de ce critère peut aussi se justifier en calculant l'espérance de R^2 et un développement limité en n montre que pour rendre le R^2 indépendant de m en première approximation il faut arriver à la définition de $R^2_{Aju}(m)$. A l'inverse des autres critères, chosir le meilleur revient à maximiser $R^2_{Aju}(m)$ et on définit ainsi :

$$\widehat{m}_{R^2_{Aju}} = \mathrm{Argmax}_{m \in \mathcal{M}} \{R^2_{Aju}(m)\}.$$

Remarque : Maximiser le critère $R^2_{Aju}(m)$ revient aussi tout simplement à minimiser $\frac{1}{n-|m|-1}\|Y-\widehat{Y}_m\|^2$ (alors que maximiser le critère R^2 revient à minimiser $\|Y-\widehat{Y}_m\|^2$).

5.4 Le critère AIC comme minimisation de la dissemblance de Kullback

Nous présentons maintenant une première minimisation assez générale de la dissemblance, valable dans un cadre non gaussien, conduisant à l'instauration d'un critère appelé AIC, puis, un "raffinement" dans le cas spécifiquement gaussien sera proposé.

Critère *AIC*

On suppose que pour tout modèle $m \in \mathcal{M}$, la distribution de Y est absolument continue par rapport à une même mesure de probabilité $\mathbb{P}$.

Le raisonnement que nous allons suivre ne se limite pas au cas du modèle linéaire (on en trouvera une version générale dans le livre de Burnham et Anderson [16] pages 230-247, ou, en moins bien fait, dans celui de Linhart et Zacchini [41], pages 242-246).

On suppose dans toute la suite que $m \in \mathcal{M}$ vérifie $m^* \subset m$.

On considère que l'on dispose d'un vecteur γ contenant les paramètres inconnus du modèle (dans notre cas, on peut se limiter à $\gamma = (\theta, \sigma^2)$ de taille $|m|+2$ en considérant que le bruit ne dépend que d'un unique paramètre, σ^2). Pour $\gamma \in \mathbb{R}^{|m|+1} \times \mathbb{R}_+$ quelconque (**pouvant être un vecteur aléatoire**, ce que nous verrons un peu plus loin), soit $g(y \mid \gamma)$ la densité de $X_{(m)}.\theta+\varepsilon$ et soit $g^* = g(. \mid \gamma^*)$ celle du vrai modèle, densité qui demeure inconnue (on considère ici que γ^* est également de taille $|m|+2$, en rajoutant des zéros pour les coordonnées correspondant à $m \setminus m^*$). Définissons :

$$\Delta(\gamma) = \mathbb{E}_{g^*}\left[\log g(Z \mid \gamma)\right],$$

où Z est un vecteur aléatoire suivant la distribution g^* du vrai modèle. On peut alors écrire la dissemblance de Kullback entre le vrai modèle et le modèle d'analyse m sous la forme :

$$K(m, m^*) = \mathbb{E}_{g^*}\left[\log g(Z \mid \gamma^*)\right] - \mathbb{E}_{g^*}\left[\Delta(\tilde{\gamma}_{(m)}(Y))\right],$$

en notant $\tilde{\gamma}_{(m)}(Y)$ le vecteur aléatoire de $\mathbb{R}^{|m|+1} \times \mathbb{R}_+$ obtenu par maximum de vraisemblance à partir du vecteur Y, réalisation (connue) d'un vecteur distribué suivant g^*.

Minimiser $K(m, m^*)$ reviendra donc à maximiser $\mathbb{E}_{g^*}\left[\Delta(\tilde{\gamma}_{(m)}(Y))\right]$.

Pour cela, on va utiliser plusieurs estimations. En premier lieu, on peut écrire, par la formule de Taylor, que pour tout vecteur aléatoire γ suffisamment proche de γ^* :

$$\log g(Z \mid \gamma) \simeq \log g(Z \mid \gamma^*) + (\gamma - \gamma^*)' \cdot \operatorname{grad} (\log g(Z \mid \gamma))_{\gamma^*} + \frac{1}{2} (\gamma - \gamma^*)' \cdot J_{\gamma^*} \cdot (\gamma - \gamma^*),$$

où J_{γ^*} est la matrice hessienne de $\log g(Z \mid \gamma)$ en γ^* (dans cette formule et dans les suivantes nous ne justifierons pas rigoureusement l'approximation des fonctions par leurs développements d'ordre 2, ce qu'en revanche on trouvera dans Linhart et Zucchini [41]). Lorsque l'on passe à l'espérance par rapport à Z qui a pour loi g^*, tout en choisissant $\gamma = \tilde{\gamma}_{(m)}(Y)$, on obtient :

$$\Delta(\tilde{\gamma}_{(m)}(Y)) \simeq \mathbb{E}_{g^*} [\log g(Z \mid \gamma^*)] + 0 + \frac{1}{2} (\tilde{\gamma}_{(m)}(Y) - \gamma^*)' \cdot \Omega^* \cdot (\tilde{\gamma}_{(m)}(Y) - \gamma^*),$$

en raison du fait que γ^* annule $\mathbb{E}_{g^*} \left[\operatorname{grad} (\log g(Z \mid \gamma))_{\gamma^*} \right]$ et en notant

$$\Omega^* = \left(\mathbb{E}_{g^*} \left[\frac{\partial^2 \log g^*(Z \mid \gamma^*)}{\partial \gamma_i \partial \gamma_j} \right] \right)_{1 \leq i,j \leq |m|+2}.$$

Ainsi, lorsque l'on passe à l'espérance de cette expression, espérance par rapport à Y qui suit également la loi g^*, on peut utiliser le fait que $\tilde{\gamma}_{(m)}(Y)$ est également l'estimateur par maximum de vraisemblance de γ^*. Or on sait que sous les hypothèses proposées,

$$\operatorname{Var} \left(\tilde{\gamma}_{(m)} \right) \sim \left(\mathbb{E}_{g^*} \left[\frac{\partial \log g(Y \mid \gamma^*)}{\partial \gamma_i} \frac{\partial \log g_{\gamma^*}(Y \mid \gamma^*)}{\partial \gamma_j} \right] \right)^{-1}_{1 \leq i,j \leq |m|+2} = (\Sigma^*)^{-1}$$

(sur les propriétés asymptotiques de l'estimateur par maximum de vraisemblance, voir par exemple les livres de Dacunha-Castelle et Duflo [20] ou de Milhaud [45]). Finalement, on obtient une première estimation de $\mathbb{E}_{g^*} \left[\Delta(\tilde{\gamma}_{(m)}(Y)) \right]$, soit :

$$\mathbb{E}_{g^*} \left[\Delta(\tilde{\gamma}_{(m)}(Y)) \right] \simeq \mathbb{E}_{g^*} [\log g(Z \mid \gamma^*)] + \frac{1}{2} \operatorname{Tr}(\Omega^* (\Sigma^*)^{-1}). \tag{9.4}$$

L'equivalence (9.4) n'est pas utilisable comme telle car il y demeure les paramètres inconnus. Une nouvelle estimation va consister à remplacer $\mathbb{E}_{g_{\gamma^*}} [\log g_{\gamma^*}(Y)]$ par $\log g_{\gamma^*}(Y)$. La loi des grands nombres permet de justifier la convergence asymptotique de cette variable aléatoire vers son espérance. On obtient maintenant que :

$$\mathbb{E}_{g^*} \left[\Delta(\tilde{\gamma}_{(m)}(Y)) \right] \simeq \log g(Y \mid \gamma^*) + \frac{1}{2} \operatorname{Tr}(\Omega^* \cdot (\Sigma^*)^{-1}).$$

Enfin, comme γ^* est inconnu, on utilise son estimation par maximum de vraisemblance, soit $\tilde{\gamma}_{(m)}$. Une deuxième formule de Taylor entraîne que :

$$\log g(Z \mid \gamma^*) \simeq \log g(Z \mid \tilde{\gamma}_{(m)}(Y)) + 0 + \frac{1}{2}(\gamma^* - \tilde{\gamma}_{(m)})' \cdot \tilde{\Omega}_{(m)} \cdot (\gamma^* - \tilde{\gamma}_{(m)}),$$

car l'estimateur du maximum de vraisemblance est obtenue par l'annulation de la dérivée par rapport à γ de $\log g(Y \mid \gamma^*)$ et avec :

$$\tilde{\Omega}_{(m)} = \left(\left[\frac{\partial^2 \log g(Z \mid \tilde{\gamma}_{(m)}(Y))}{\partial \gamma_i \partial \gamma_j} \right] \right)_{1 \leq i,j \leq |m|+2}.$$

Par suite, en utilisant le fait que $\tilde{\Omega}_{(m)}$ converge asymptotiquement quand $n \to \infty$ vers Ω^* (car on a supposé que $\tilde{\gamma}_{(m)}(Y)$ converge en probabilité vers γ^*), et en remplaçant ce qui précède dans (9.5), on obtient que :

$$\mathbb{E}_{g^*} \left[\Delta(\tilde{\gamma}_{(m)})(Y) \right] \simeq \log g(Z \mid \tilde{\gamma}_{(m)}(Y)) + \mathrm{Tr}(\Omega^* (\Sigma^*)^{-1}). \tag{9.5}$$

Une dernière inconnue demeure : $\mathrm{Tr}(\Omega^* \cdot (\Sigma^*)^{-1})$. On montre facilement que

$$\mathbb{E}_{g^*} \left[\frac{\partial \log g_{\gamma^*}(Y)}{\partial \gamma_i} \frac{\partial \log g_{\gamma^*}(Y)}{\partial \gamma_j} \right] = -\mathbb{E}_{g_{\gamma^*}} \left[\frac{\partial^2 \log g_{\gamma^*}(Y)}{\partial \gamma_i \partial \gamma_j} \right]$$

car g^* est une densité de probabilité dont le support ne dépend pas des paramètres, et donc $\displaystyle\int_{\mathbb{R}} \frac{\partial^2 g(x \mid \gamma)}{\partial \gamma_i \partial \gamma_j} dx = 0$. En conséquence, et sous les hypothèses faites,

$$\mathrm{Tr}(\Omega^* (\Sigma^*)^{-1}) = -\big(|m| + 2\big). \tag{9.6}$$

On en arrive ainsi en utilisant (9.4), (9.5) et (9.6) à montrer que **minimiser la dissemblance de Kullback revient (après ces nombreuses approximations) à maximiser :**

$$\log g(Z \mid \tilde{\gamma}_{(m)}(Y)) - \big(|m| + 2\big).$$

On peut, en multipliant le tout par -2, écrire que la minimisation la dissemblance de Kullback revient à celle du critère appelé *AIC* (pour Akaiké Information Criterium, introduit par Akaiké en 1973 [1]), tel que

$$\begin{aligned} AIC(m) &= -2 \log g(Z \mid \tilde{\gamma}_{(m)}(Y)) + 2\big(|m| + 2\big) \\ &= -2 \times \log(\text{Vraisemblance maximisée}) + 2 \times \text{ Nombre de paramètres.} \end{aligned}$$

Suivant ce critère, on choisira m tel que

$$\widehat{m}_{AIC} = \mathrm{Argmin}_{m \in \mathcal{M}} \{AIC(m)\}.$$

Remarque : 1/ nous avons présenté la construction du critère AIC dans un cas général pour lequel l'estimateur des paramètres est l'estimateur par maximum de vraisemblance. Dans le cas du modèle linéaire, l'estimation par maximum de vraisemblance correspond à l'estimation par moindres carrés dans le cas de distributions gaussiennes.
2/ Dans le cas gaussien, le critère AIC pourra s'écrire (à des constantes près, non importantes pour sa minimisation) :

$$AIC(m) = n \log\left(\tilde{\sigma}^2_{(m)}\right) + 2\left(|m| + 2\right),$$

puisqu'alors log(Vraisemblance maximisée) $= -n \log \tilde{\sigma}_{(m)} - \dfrac{n}{2} \log 2\pi - \dfrac{n}{2}$.

Critère AIC_c

Dans le cas spécifiquement gaussien, on peut être plus précis dans les estimations effectuées précédemment. Soit $m \in \mathcal{M}$. Alors, la dissemblance de Kullback entre la loi de l'estimation issue de ce modèle et le vrai modèle est :

$$K(m, m^*) = \mathbb{E}_{m^*}\left[\frac{n}{2}\left(\log\left(\frac{\tilde{\sigma}^2_{(m)}}{\sigma^2_*}\right) + \frac{\sigma^2_*}{\tilde{\sigma}_{(m)}} - 1\right) + \frac{1}{2\tilde{\sigma}_{(m)}}||\mu^* - \widehat{Y}_{(m)}||^2\right].$$

Une écriture plus explicite de cette quantité n'est pas possible, mais une première simplification pourra être faite si l'**on fait l'hypothèse que** $m^* \in m$. Alors, en utilisant pleinement la distribution gaussienne du bruit :

$$\tilde{\sigma}^2_{(m)} = \frac{1}{n}||Y - \widehat{Y}_{(m)}||^2 = \frac{1}{n}||P_{(m)^\perp}\varepsilon^*||^2 \sim \frac{\sigma^2_*}{n}\,\chi^2(n - (|m| + 1))$$

d'après le Théorème de Cochran et en appellant $P_{(m)^\perp}$ l'opérateur de projection orthogonale sur $[X_{(m)}]^\perp$. De même,

$$||\mu^* - \widehat{Y}_{(m)}||^2 = ||P_{(m)}\varepsilon^*||^2 \sim \sigma^2_* \cdot \chi^2((|m| + 1))$$

et d'après ce même Théorème de Cochran,

$$||\mu^* - \widehat{Y}_{(m)}||^2 \text{ est indépendant de } \tilde{\sigma}^2_{(m)}.$$

Par suite,

$$K(m, m^*) = \frac{n}{2}\left(\mathbb{E}_{m^*}\left[\log\left(\tilde{\sigma}^2_{(m)}\right)\right] - \log\left(\sigma^2_*\right) - 1 + \frac{n}{(n - |m| - 3)} + \frac{|m| + 1}{(n - |m| - 3)}\right),$$

d'après l'expression de l'espérance d'une loi de Fisher. Il nous reste à déterminer l'expression de $\mathbb{E}_{m^*}\left[\log\left(\tilde{\sigma}^2_{(m)}\right)\right]$. Comme on a supposé que $m^* \subset m$, un tel calcul est possible, mais le résultat dépendra de $\log(\sigma^2_*)$. On va donc préférer utiliser un estimateur de cette espérance, à savoir $\log\left(\tilde{\sigma}^2_{(m)}\right)$, qui converge vers $\mathbb{E}_{m^*}\left[\log\left(\tilde{\sigma}^2_{(m)}\right)\right]$ d'après la loi des grands nombres (la fonction log étant continue sur $\mathbb{R}^*_+$) puisque $m^* \subset m$.

En enlevant les parties constantes et en multipliant le tout par 2, on voit que minimiser $K(m, m^*)$ revient approximativement à minimiser le critère suivant, noté AIC_c (pour AIC corrigé), établi en 1989 par Hurvich et Tsai [34], et qui s'écrit pour $m \in \mathcal{M}$,

$$AIC_c(m) = n \log\left(\tilde{\sigma}^2_{(m)}\right) + n\,\frac{n + |m| + 1}{n - |m| - 3}.$$

Le modèle sélectionné sera donc le modèle $\widehat{m}$ tel que :

$$\widehat{m}_{AIC_c} = \mathrm{Argmin}_{m\in\mathcal{M}}\{AIC_c(m)\}.$$

Lorsque $n \to \infty$ (et avec m fixé), on peut trouver un équivalent au critère AIC_c, puisque :

$$\frac{n + |m| + 1}{n - |m| - 3} \sim 1 + 2\,\frac{|m| + 2}{n}.$$

Cela reviendra donc asymptotiquement et approximativement à minimiser le critère AIC, tel que :

$$AIC(m) = n \log\left(\tilde{\sigma}^2_{(m)}\right) + 2\big(|m| + 2\big).$$

Cependant, ce critère donne de mauvais résultats pour n petit, et on préférera donc, dans le cas gaussien, utiliser AIC_c pour tout n.

Plus généralement, observons que la sélection de modèle se fait en utilisant un critère de type vraisemblance pénalisée, c'est-à-dire que pour un modèle $m \in \mathcal{M}$, il s'écrira sous la forme :

$$\mathrm{Crit}(m) = -f(\text{Vraisemblance maximisée}) + \mathrm{Pen}(|m|),$$

où la fonction f est une fonction croissante (par exemple l'identité ou le log), la vraisemblance dépend par exemple de $\tilde{\sigma}^2_{(m)}$ et décroît lorsque $|m|$ croît (plus le modèle est important mieux il approche les données et il est donc plus "vraisemblable"), et la pénalisation Pen est une fonction croissante de $|m|$.

5.5 Le critère BIC comme une extension du modèle AIC

Le critère BIC (Bayesian Information Criterium) introduit en 1978 (voir par exemple Akaiké [2] ou Schwarz [53]), est extension de l'écriture générale du critère AIC, puisque l'on posera par définition :

$$BIC(m) = n \log\left(\tilde{\sigma}^2_{(m)}\right) + \log n \left(|m| + 2\right).$$

Ce critère (que l'on va également minimiser) est obtenu à partir d'a priori sur le modèle et à partir d'une décomposition de sa vraisemblance. Nous n'entrerons pas plus dans les détails, car, si ce critère donne de très bons résultats, son écriture n'est pas forcément très bien justifiée. Le terme en $\log n$ a été remplacé par $\log\log n$ par Hannan et Quine [32], et par une suite (c_n) telle que $(\log\log n)^{-1} c_n \to \infty$ et $n^{-1} c_n \to 0$ quand $n \to \infty$ par Rao et Wu [50].

Conclusion : Au final, dans le cadre du modèle linéaire, nous disposons de 6 critères à minimiser et un critère à maximiser R^2_{Aju} pour sélectionner un "bon" modèle prédictif :

$$\begin{aligned}
Cp(m) &= \frac{\|Y - \widehat{Y}_m\|^2}{\|Y - \widehat{Y}_{(m_k)}\|^2} + 2\,\frac{(|m|+1)}{n} \\
PRESS(m) &= \sum_{i=1}^{n} \left(Y_i - (X_{(m)}\,\widehat{\theta}^{-i}_{(m)})_i\right)^2 \\
R^2_{Aju}(m) &= 1 - \frac{n-1}{n-|m|-1}\,\frac{\|Y - \widehat{Y}_{(m)}\|^2}{\|Y - \overline{Y}\|^2} \\
AIC(m) &= n \log\left(\|Y - \widehat{Y}_{(m)}\|^2\right) + 2(|m|+1) \\
AIC_c(m) &= n \log\left(\|Y - \widehat{Y}_{(m)}\|^2\right) + n\,\frac{n+|m|+1}{n-|m|-3} \\
BIC(m) &= n \log\left(\|Y - \widehat{Y}_{(m)}\|^2\right) + \log n\,(|m|+1).
\end{aligned}$$

Nous allons étudier 5 de ces 6 critères, le critère $PRESS$ étant de nature un peu différente, nous n'en dirons que quelques mots plus loin.

6 Probabilité de préférer un modèle à un autre

On suppose deux modèles m_1 et m_2 appartenant à $\mathcal{M}$. En notant Crit le critère utilisé parmi 5 des critères précédents (le critère $PRESS$ n'est pas considéré ici car demandant une étude spécifique). On s'intéresse ici à l'évènement $\text{Crit}(m_1) \leq \text{Crit}(m_2)$, qui signifie que pour les critères Cp, AIC, AIC_c ou BIC on choisira m_1 plutôt que m_2, l'inverse pour le critère R^2_{Aju}.

On remarque tout d'abord que **si** $|m_1| = |m_2|$ **alors ces** 5 **critères conduisent à préférer le modèle le mieux ajusté, c'est-à-dire le modèle ayant une somme des carrés des résidus la plus faible ; c'est aussi ce que choisit le critère** R^2**...**

En fixant désormais $|m_1| < |m_2|$, nous allons faire apparaître la statistique de type Fisher suivante :

$$\widehat{F}(m_1, m_2) := \Big(\frac{n - |m_2| - 1}{|m_2| - |m_1|}\Big) \frac{\|Y - \widehat{Y}_{(m_1)}\|^2 - \|Y - \widehat{Y}_{(m_2)}\|^2}{\|Y - \widehat{Y}_{(m_2)}\|^2}.$$

Commençons à étudier le cas du critère Cp qui est un peu différent des 4 autres :

$$\begin{aligned} \mathbb{P}\left[\mathrm{Cp}(m_1) \leq \mathrm{Cp}(m_2)\right] &= \mathbb{P}\Big[\|Y - \widehat{Y}_{(m_1)}\|^2 - \|Y - \widehat{Y}_{(m_2)}\|^2 \leq 2\,\frac{|m_2| - |m_1|}{n}\,\|Y - \widehat{Y}_{(m_k)}\|^2\Big] \\ &= \mathbb{P}\Big[\widehat{F}(m_1, m_2) \leq 2\,\frac{n - |m_2| - 1}{n}\,\frac{\|Y - \widehat{Y}_{(m_k)}\|^2}{\|Y - \widehat{Y}_{(m_2)}\|^2}\Big] \end{aligned} \tag{9.7}$$

Pour le critère R^2_{Aju} des calculs similaires entraînent que :

$$\mathbb{P}\left[R^2_{Aju}(m_1) \leq R^2_{Aju}(m_2)\right] = \mathbb{P}\Big[\widehat{F}(m_1, m_2) \geq 1\Big]. \tag{9.8}$$

Le même genre de calculs pour les critères AIC et BIC (pour AIC_c nous aurons asymptotiquement les mêmes résultats qu'avec le critère AIC) conduisent à :

$$\begin{aligned} &\mathbb{P}\left[\mathrm{AIC}(m_1) \leq \mathrm{AIC}(m_2)\right] \\ &\quad = \mathbb{P}\Big[\widehat{F}(m_1, m_2) \leq \Big(\frac{n - |m_2| - 1}{|m_2| - |m_1|}\Big)\Big(\exp\big(\frac{2(|m_2| - |m_1|)}{n}\big) - 1\Big)\Big] \end{aligned} \tag{9.9}$$

$$\begin{aligned} &\mathbb{P}\left[\mathrm{BIC}(m_1) \leq \mathrm{BIC}(m_2)\right] \\ &\quad = \mathbb{P}\Big[\widehat{F}(m_1, m_2) \leq \Big(\frac{n - |m_2| - 1}{|m_2| - |m_1|}\Big)\Big(\exp\big((|m_2| - |m_1|)\,\frac{\log n}{n}\big) - 1\Big)\Big]. \end{aligned} \tag{9.10}$$

De ces différentes probabilités on va pouvoir déduire les probabilités asymptotiques de choisir un autre modèle que m^*, mais il nous faut distinguer maintenant le cas des modèles sur-ajustés et celui des faux modèles.

• Probabilité de préférer un sur-modèle

Ici on suppose que $m_1 = m^*$, le vrai modèle, et $m_1 \subset m_2$ avec $m_2 \neq m_1$. Si le critère choisit le modèle m_2 plutôt que le vrai modèle $m_1 = m^*$, le critère a tendance à surajuster. Nous sommes ici exactement dans le cadre d'un test d'hypothèse de l'hypothèse H_0 : "le sous-modèle m^* est le vrai modèle" contre H_1 : "le vrai modèle est le modèle m_2" et la statistique est celle que nous avons déjà utilisée dans le chapitre 3 pour ce problème de test. En utilisant les résultats asymptotiques du chapitre 8, on en arrive à la proposition suivante :

Proposition 9.1 *Soit la famille de modèles linéaires statistiques (8.3) sous l'hypothèse* **H**, *c'est-à-dire que l'on suppose les postulats* **P1-3**, *plus l'équidistribution des erreurs mais pas forcément leur gaussianité (juste l'existence d'un moment d'ordre* 2*). Si* m_2 *est un modèle contenant strictement* $m_1 = m^*$, *on a lorsque* $n \to \infty$ *:*

$$\begin{aligned}
\mathbb{P}\left[Cp(m_1) \geq Cp(m_2)\right] &\underset{n\to+\infty}{\longrightarrow} \mathbb{P}\left[\chi^2(|m_2|-|m^*|) \geq 2(|m_2|-|m^*|)\right] \\
\mathbb{P}\left[R^2_{Aju}(m_1) \leq R^2_{Aju}(m_2)\right] &\underset{n\to+\infty}{\longrightarrow} \mathbb{P}\left[\chi^2(|m_2|-|m^*|) \geq (|m_2|-|m^*|)\right] \\
\mathbb{P}\left[AIC(m_1) \geq AIC(m_2)\right] &\underset{n\to+\infty}{\longrightarrow} \mathbb{P}\left[\chi^2(|m_2|-|m^*|) \geq 2(|m_2|-|m^*|)\right] \\
\mathbb{P}\left[BIC(m_1) \geq BIC(m_2)\right] &\underset{n\to+\infty}{\longrightarrow} 0.
\end{aligned}$$

Démonstration : Ceci se montre en utilisant le résultat de le Théorème 8.3 du Chapitre 8 qui montre que sous les hypothèses, $\widehat{F}(m_1,m_2) \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \frac{1}{|m_2|-|m^*|}\,\chi^2(|m_2|-|m^*|)$. De plus $\frac{\|Y-\widehat{Y}_{(m_k)}\|^2}{\|Y-\widehat{Y}_{(m_2)}\|^2} \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} 1$. Avec l'aide de développements limités on obtient bien les résultats de la Proposition 9.1 à partir de (9.7), (9.8) et (9.9).
De plus pour tout $C>0$, il existe n_0 tel que pour tout $n \geq n_0$, $\left(\frac{n-|m_2|-1}{|m_2|-|m_1|}\right)\left(\exp\left((|m_2|-|m_1|)\,\frac{\log n}{n}\right)-1\right) \simeq \log n > C$. Ainsi pour $n \geq n_0$, en utilisant (9.10), on a :

$$\mathbb{P}\left[\widehat{F}(m_1,m_2) \geq \left(\frac{n-|m_2|-1}{|m_2|-|m_1|}\right)\left(\exp\left((|m_2|-|m_1|)\,\frac{\log n}{n}\right)-1\right)\right] \leq \mathbb{P}\left[\widehat{F}(m_1,m_2) \geq C\right].$$

Comme $\widehat{F}(m_1,m_2) \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \frac{\chi^2(|m_2|-|m^*|)}{|m_2|-|m^*|}$, on a

$$\mathbb{P}\left[\mathrm{BIC}(m_1) \geq \mathrm{BIC}(m_2)\right] \leq \mathbb{P}\left[\widehat{F}(m_1,m_2) \geq C\right] \underset{n\to+\infty}{\longrightarrow} \mathbb{P}\left[\frac{\chi^2(|m_2|-|m^*|)}{|m_2|-|m^*|} \geq C\right] \leq \frac{1}{C},$$

d'après l'Inégalité de Markov. Mais comme ceci est vrai pour tout $C>0$, on obtient bien le résultat. ■

Conclusion : Que constate-t-on à partir de cette proposition ? Que **le seul critère évitant asymptotiquement de surajuster est le BIC,** les trois autres (ou quatre autres si on considere le critère AIC_c qui a les mêmes propriétés asymptotiques que le critère AIC) ayant une probabilité positive de surajuster. En effet, la probabilité asymptotique de sur-ajustement prend en compte tous les modèles possibles m_2 contenant strictement m^* (qui restent cependant en nombre fini $\leq 2^k$ puisque l'on suppose que k reste fixé), tend également vers 0 pour le critère BIC et est strictement positive pour les autres critères.
On remarque également que **le critère R^2_{Aju} a encore plus tendance asymptotiquement à sur-ajuster que les critères Cp ou AIC.**

Remarque : Ce résultat concernant le BIC serait toujours valable si on remplaçait le $\log n$ dans l'expression du BIC par une autre vitesse c_n telle que $c_n \underset{n\to+\infty}{\longrightarrow} \infty$ (on peut appeler un tel critère GIC, comme Generalized Information Theory). Soit :

$$GIC(m) \quad = \quad n \log\big(\|Y - \widehat{Y}_{(m)}\|^2\big) + c_n\,|m|. \tag{9.11}$$

On ne peut tout de même pas choisir n'importe quelle suite (c_n) puisque le développement limité utilisé n'est valable que si $c_n = o(n)$.

• Probabilité de préférer un faux-modèle

Considérons maintenant un faux modèle m_1 et définissons $m_2 = m_1 \cup m^*$. Ainsi, le modèle m_2 contient le vrai modèle mais également toutes les variables explicatives de m_1 : m_2 est donc soit le vrai modèle (au cas où m_1 est contenu strictement dans m^*, donc dans le cas de sous-ajustement), soit un modèle sur-ajusté. On peut alors estimer la probabilité asymptotique de choisir m_1 plutôt que m_2. Pour ce faire on remarque que sous l'hypothèse **H**, $\|Y-\widehat{Y}_{(m_2)}\|^2$ suit asymptotiquement une loi $\sigma^2\,\chi^2(n-|m_2|-1)$ car m_2 contient m^* et $\|Y - \widehat{Y}_{(m_1)}\|^2 = \big\|P_{[X^{(m_1)}]^\perp}\varepsilon + P_{[X^{(m_1)}]^\perp}X^{(m^*)}\theta^{(m^*)}\big\|^2$, ce qui signifie que $\|Y-\widehat{Y}_{(m_1)}\|^2$ suit une loi du Chi-deux décentrée. En utilisant le Théorème de Pythagore comme dans le chapitre 3 et le comportement asymptotique donné au Theorème 8.3 du chapitre 8 on en déduit que sous l'hypothèse **H** :

$$\widehat{F}(m_1, m_2) \underset{n\to\infty}{\overset{\mathcal{L}}{\sim}} \big(|m_2| - |m_1|\big)\,\chi^2_{NC}\big(t(n, m_1)\,,\,|m_2| - |m_1|\big),$$

où $\chi^2_{NC}(t, d)$ est une loi du χ^2 décentré de paramètre de décentrage t et de degré de liberté d et

$$t(n, m_1) := P_{[X^{(m_1)}]^\perp}X^{(m^*)}\theta^{(m^*)}.$$

Il est tout à fait possible que $t(n, m_1)$ ne converge pas lorsque $n \to \infty$ (voir des exemples un peu plus loin). Il n'est donc pas possible de donner un résultat qui ne dépende pas du comportement asymptotique de $t(n, m_1)$ et donc de celui de X. On peut cependant déjà énoncer le résultat suivant :

Proposition 9.2 *Soit la famille de modèles linéaires statistiques (8.3) sous l'hypothèse les postulats* **P1-4**. *Si m_1 est un faux modèle et si $m_2 = m_1 \cup m^*$, si $t(n, m_1) \geq 2$ et lorsque n est suffisamment grand :*

$$\mathbb{P}\left[\widehat{F}(m_1, m_2) \leq \frac{1}{8}\,\frac{t(n, m_1)}{|m_2| - |m_1|}\right] \leq \exp\Big(-\frac{1}{8}\,\frac{t(n, m_1)}{|m_2| - |m_1|}\Big).$$

Démonstration : Comme on a choisi m_1 un sous-modèle de m_2, sous les postulats **P1-4**, on peut utiliser les résultats asymptotiques du chapitre 8 concernant $\widehat{F}(m_1, m_2)$. On

peut écrire que pour n suffisamment grand, comme le dénominateur de $\frac{\sigma^2}{\sigma^2}\widehat{F}(m_1,m_2)$ tend presque-sûrement vers 1 (on a donc tout renormalisé par σ^2),

$$\begin{aligned}
\mathbb{P}\left[\widehat{F}(m_1,m_2) \leq \frac{1}{8}\,\frac{t(n,m_1)}{|m_2|-|m_1|}\right] &\leq \mathbb{P}\left[\chi^2_{NC}\big(t(n,m_1)\,,\,|m_2|-|m_1|\big) \leq \frac{1}{4}\,t(n,m_1)\right] \\
&\leq \mathbb{P}\left[\chi_{NC}\big(t(n,m_1)\,,\,|m_2|-|m_1|\big) \leq \frac{1}{2}\sqrt{t(n,m_1)}\right] \\
&\leq \mathbb{P}\left[\chi\big(|m_2|-|m_1|\big) \geq \sqrt{t(n,m_1)} - \frac{1}{2}\,\sqrt{t(n,m_1)}\right] \\
&\leq \mathbb{P}\left[\chi^2\big(|m_2|-|m_1|\big) \geq \frac{1}{4}\,t(n,m_1)\right] \\
&\leq \exp\Big(-\frac{1}{8}\,\frac{t(n,m_1)}{|m_2|-|m_1|}\Big),
\end{aligned}$$

où la dernière inégalité est liée au résultat suivant :

Lemme 9.1 *Soit $C \geq 1/2$. Alors pour tout $d \in \mathbb{N}^*$,*

$$P\Big[\chi^2(d) \geq C\,d\Big] \leq 2\,\exp(-C/2).$$

<u>*Démonstration*</u> : Ceci se montre facilement dans le cas $d \geq 2$ en utilisant l'inégalité de Bienaymé-Tchebytchev, puisque

$$\begin{aligned}
\mathbb{P}\left[\chi^2(d) \geq C\,d\right] &= \mathbb{P}\left[\exp(\chi^2(d)/2d) \geq \exp(C/2)\right] \\
&\leq \mathbb{E}\left[\exp(\chi^2(d)/2d)\right]\exp(-C/2) \\
&\leq \big(1-1/d\big)^{-d/2}\exp(-C/2),
\end{aligned}$$

et $\big(1-1/d\big)^{-d/2} \leq 2$ pour tout $d \geq 2$. Dans le cas $d = 1$, on écrit que :

$$\begin{aligned}
\mathbb{P}\left[\chi^2(1) \geq C\right] &= \frac{1}{\sqrt{2}\,\Gamma(1/2)}\int_C^\infty t^{-1/2}e^{-t/2}dt \\
&= \frac{1}{\sqrt{2}\,\Gamma(1/2)}\Big(\big[-2t^{-1/2}e^{-t/2}\big]_C^\infty - \frac{1}{2}\int_C^\infty t^{-3/2}e^{-t/2}dt\Big) \\
&\leq \frac{1}{\sqrt{2}\,\Gamma(1/2)}\big(2C^{-1/2}e^{-C/2}\big) \leq 2e^{-C/2},
\end{aligned}$$

dès que $C \geq 1/2$ (on a également utilisé une intégration par parties à la deuxième étape). ■

On peut également obtenir des résultats dans le cas de l'hypothèse **H**, c'est-à-dire dans le cas où les postulats **P1-3** sont vérifiés, mais pas nécessairement **P4** (le caractère gaussien de l'erreur). Ainsi obtient-on :

Proposition 9.3 *Soit la famille de modèles linéaires statistiques (8.3) sous l'hypothèse* **H**. *Soit* $C \geq 1/2$ *fixé. Si* m_1 *est un faux modèle, si* $m_2 = m_1 \cup m^*$ *et si* $\frac{1}{8}\,\frac{t(n,m_1)}{|m_2|-|m_1|} \geq C$ *pour tout* $n \geq n_0$ *avec* $n_0 \in N$, *alors lorsque* $n \geq n_0$:

$$\mathbb{P}\left[\widehat{F}(m_1,m_2) \leq C\right] \leq \frac{1}{2C}.$$

Démonstration : Comme précédemment, la convergence presque-sûre vers 1 du dénominateur de $\widehat{F}(m_1,m_2)$ renormalisé par σ^2 ayant également lieu sous l'hypothèse **H**, on peut écrire que pour n suffisamment grand,

$$\begin{aligned}
\mathbb{P}\left[\widehat{F}(m_1,m_2) \leq C\right] &\leq \mathbb{P}\left[\left\|t(n,m_1)+W_n\right\|^2 - \leq \frac{1}{4}\,t(n,m_1)\right] \\
&\leq \mathbb{P}\left[\xi_n \geq \frac{1}{4}\,t(n,m_1)\right] \\
&\leq \mathbb{P}\left[\frac{\xi_n}{|m_2|-|m_1|} \geq 2\,C\right]
\end{aligned}$$

avec $W_n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}_{|m_2|-|m_1|}(0, I_{|m_2|-|m_1|})$ et $\xi_n \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \chi^2(|m_2|-|m_1|)$. On utilise alors la convergence en loi et l'Inegalité de Markov pour obtenir le résultat. ■

Il ne nous reste plus qu'à trouver des conditions telles que $t(n,m_1) \to \infty$, car alors $\widehat{F}(m_1,m_2) \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \infty$ d'après la Proposition 9.2. Nous donnons ici une condition qui permettra de montrer que les différents critères présentés ne choisissent pas asymptotiquement un faux-modèle :

Proposition 9.4 *Soit la famille de modèles linéaires statistiques (8.3) sous l'hypothèse* **H**. *On suppose de plus que* $X_{(m_k)}$ *vérifie*

$$\frac{1}{d(n)}\,X'_{(m_k)}\,X_{(m_k)} \underset{n\to+\infty}{\longrightarrow} M, \tag{9.12}$$

où $d(n)\,\log^{-1}(n) \underset{n\to+\infty}{\longrightarrow} \infty$ *et* M *est une matrice définie positive de taille* k. *Alors :*

$$\mathbb{P}\left[\text{“Crit choisit un faux-modèle”}\right] \underset{n\to+\infty}{\longrightarrow} 0,$$

où Crit=Cp, AIC, AIC$_c$ ou BIC.

Démonstration : Comme le nombre de faux-modèle est fini, il suffit de démontrer que la probabilité de préférer un seul faux-modèle quelconque tend vers 0. Soit donc m_1 un faux modèle et avec $m_2 = m_1 \cup m^*$, on a :

$$t(n,m_1) = \left\|P_{[X^{(m_1)}]^\perp} X^{(m_k)}\theta^{(m_k)}\right\|^2 = \left\|X^{(m_k)}\theta^{(m_k)}\right\|^2 - \left\|P_{[X^{(m_1)}]} X^{(m_k)}\theta^{(m_k)}\right\|^2$$

par le Théorème de Pythagore. Tout d'abord on a d'après les hypothèses de la proposition, lorsque $n \to \infty$,

$$\left\|X^{(m_k)}\theta^{(m_k)}\right\|^2 \simeq d(n)\,\theta'^{(m_k)}\,M\,\theta^{(m_k)}.$$

Ensuite, toujours lorsque $n \to \infty$,

$$\begin{aligned}\left\|P_{[X^{(m_1)}]}X^{(m_k)}\theta^{(m_k)}\right\|^2 &\simeq d(n)\,\theta'^{(m_k)}(X^{(m_k)})'X^{(m_1)}\left((X^{(m_1)})'X^{(m_1)}\right)^{-1}(X^{(m_1)})'(X^{(m_k)})\,\theta^{(m_k)}\\ &\simeq d(n)\,\theta'^{(m_k)}M'_{m_k,m_1}\,M^{-1}_{m_1,m_1}M_{m_k,m_1}\,\theta^{(m_k)},\end{aligned}$$

où M_{m_k,m_1} est la matrice extraite de M en choisissant toutes les lignes et les colonnes définies par le modèle m_1 et M_{m_1,m_1} l'extraction de M en choisissant les lignes et colonnes appartenant à m_1. Or M peut être définie comme la matrice de Gram (matrice des produits scalaires) d'une famille libre de k vecteurs V_1, $V_2, \cdots, V_k$ de $\mathbb{R}^k$, car M est supposée être une matrice définie positive. Alors $M'_{m_k,m_1}\,M^{-1}_{m_1,m_1}M_{m_k,m_1}$ est la matrice associée à la forme quadratique $(y_1, \cdots, y_k) \to \left\|P_{[X^{(m_1)}]}\sum_{i=1}^k y_iV_i\right\|^2$ et ainsi par le même raisonnement :

$$t(n, m_1) \simeq d(n)\left\|P_{[X^{(m_1)}]^\perp}\sum_{i=1}^k z_iV_i\right\|^2,$$

où les z_i sont les coordonnées de $\theta^{(m_k)}$ quand n est grand. Comme m_1 est un faux modèle, $\theta^{(m_k)}$ a ses coordonnées qui n'appartiennent pas à $[X^{(m_1)}]$, donc $\sum_{i=1}^k z_iV_i$ n'appartient pas à $[X^{(m_1)}]$, donc $P_{[X^{(m_1)}]^\perp}\sum_{i=1}^k z_iV_i = C_1(m_1)$. On obtient donc que

$$t(n, m_1) \simeq C(m_1)\,d(n)$$

lorsque $n \to \infty$ avec $C(m_1) > 0$ ne dépendant pas de n.
On peut alors utiliser la Proposition 9.3 et ainsi, pour tout $C > 0$ et n suffisamment grand tel que $\frac{1}{8}\,\frac{C(m_1)}{|m_2|-|m_1|}\,d_n \geq C$,

$$\mathbb{P}\left[\widehat{F}(m_1, m_2) \leq C\right] \leq \frac{1}{2C}.$$

On peut revenir maintenant aux 5 critères que nous avons étudiés et utiliser les résultats asymptotiques de la Proposition 9.1. Ainsi si "Crit"=Cp, AIC ou AIC_c, et avec n suffisamment grand,

$$\mathbb{P}\left[\text{Crit}(m_1) \leq \text{Crit}(m_2)\right] \simeq \mathbb{P}\left[\widehat{F}(m_1, m_2) \leq 2\right] \underset{n\to+\infty}{\longrightarrow} 0.$$

On a le même genre de résultat pour le critère R^2_{Aju}. Pour le critère BIC, pour n suffisamment grand,

$$\mathbb{P}\left[\text{BIC}(m_1) \leq \text{BIC}(m_2)\right] \simeq \mathbb{P}\left[\widehat{F}(m_1, m_2) \leq \log n\right] \underset{n\to+\infty}{\longrightarrow} 0.$$

■

Remarque : Sous les postulats **P1-4**, c'est-à-dire sous l'hypothèse gaussienne,

$$\mathbb{P}\left[\text{Crit}(m_1) \leq \text{Crit}(m_2)\right] \leq 2\,\exp\Big(-\frac{1}{8}\,\frac{C(m_1)}{|m_2|-|m_1|}\,d_n\Big).$$

En chosissant $d_n >> \log(n)$,

$$\mathbb{P}\left[\text{BIC}(m_1) \leq \text{BIC}(m_2)\right] \simeq \mathbb{P}\left[\widehat{F}(m_1,m_2) \leq \log n\right] \leq 2\,\exp\Big(-\frac{1}{8}\,\frac{C(m_1)}{|m_2|-|m_1|}\,d_n\Big),$$

La probabilité de choisir un faux-modèle tend donc en général très vite vers 0.

On remarque ici que contrairement à la probabilité de sur-ajusté, le critère BIC est le "moins bon" pour ne pas choisir asymptotiquement un faux-modèle. Mais la différence est le plus souvent totalement négligeable car d_n est en général de l'ordre de n ou tout au moins d'une puissance positive de n. On en arrive à la principale conclusion de ce chapitre :

Conclusion : Pour choisir un modèle nous recommandons d'**utiliser le critère BIC car il est le seul parmi les** 6 **critères proposés à asymptotiquement choisir le "vrai" modèle avec une probabilité tendant vers** 1**. Les** 5 **autres critères (Cp, PRESS,** R^2_{Aju}**, AIC, AIC**$_c$**) ont toujours une probabilité positive de surajuster** (il a été montré par exemple dans Li, 1987, que le critère PRESS possède le même comportement asymptotique que les critères Cp, AIC ou AIC$_c$, tout en étant bien plus lourd numériquement). Le critère R^2_{Aju} **est le critère le moins intéressant et nous ne recommandons pas son utilisation**. Enfin ces propos concernant le critère BIC peuvent être généralisés au critère GIC défini en (9.11), à condition que la vitesse c_n du critère GIC soit telle que $c_n >> d_n$ avec $c_n = o(n)$.

Il est intéressant de connaître également le comportement asymptotique du risque quadratique quand on utilise un modèle sélectionné suivant un des critères précédents, car cela mesure en quelque sorte l'erreur commise sur l'estimation du modèle. On peut tout d'abord définir le risque quadratique qu'un modèle fixé, disons $m \in \mathcal{M}$, soit le modèle sélectionné, c'est-à-dire

$$R_n(m) = \mathbb{E}_{\,m^*}\left(\|\mu^* - \widehat{Y}_{(m)}\|^2 \cdot \mathbb{1}_{\{\widehat{m}=m\}}\right).$$

Le risque quadratique du modèle sélectionné est :

$$R_n = \mathbb{E}_{\,m^*}\left(\|\mu^* - \widehat{Y}_{(\widehat{m})}\|^2\right).$$

Il est clair que :

$$R_n = \sum_{m \in \mathcal{M}} R_n(m). \tag{9.13}$$

Définissons également :

$$p_n(m) = \mathbb{P}\left(\widehat{m} = m\right).$$

On peut alors montrer la double propriété suivante :

Propriété 9.1 *Sous les hypothèses de la Proposition 9.4, lorsque le critère utilisé est le critère GIC avec $c_n = o(n)$ et $c_n >> d_n$, et si $\mathbb{E}\,\varepsilon_0^4 < \infty$ (où (ε_i) est la suite des erreurs du modèle) alors $R_n \underset{n\to+\infty}{\longrightarrow} R_n(m^*) = |m^*|\,\sigma_*^2$.*

<u>*Démonstration*</u> : On peut encore écrire que $R_n(m) = \mathbb{E}_{m^*}\left(\|\mu^* - \widehat{Y}_{(m)}\|^2 \cdot \mathbb{1}_{\{\widehat{m}=m\}}\right)$, soit :

$$\begin{aligned} R_n(m) &= \mathbb{E}_{m^*}\left(\|\mu^* - \mu^*_{(m)}\|^2\,\mathbb{1}_{\{\widehat{m}=m\}} + \|\mu^*_{(m)} - \widehat{Y}_{(m)}\|^2\,\mathbb{1}_{\{\widehat{m}=m\}}\right) \\ &= \|P_{[X^{(m)}]^\perp} X^{(m_k)}\theta^{(m_k)}\|^2\, p_n(m) + \mathbb{E}_{m^*}\left(\|P_{[X^{(m)}]}\varepsilon\|^2\,\mathbb{1}_{\{\widehat{m}=m\}}\right) \\ &= J_1(m) + J_2(m). \end{aligned}$$

Concernant $J_1(m)$, si m est un sur-modèle ou le vrai modèle m^* alors $J_1(m) = 0$ car $P_{[X^{(m)}]^\perp} X^{(m_k)}\theta^{(m_k)} = 0$.
Sinon, on peut écrire que pour un faux-modèle m et avec les notations précédentes, pour n suffisamment grand,

$$J_1(m) = \|t(n, m_1)\|^2\, p_n(m) \quad \leq \quad \leq 2C(m_1)\, d_n\, \exp\Big(-\frac{1}{8}\,\frac{C(m_1)}{|m_2| - |m_1|}\, d_n\Big)$$

donc $J_1(m) \underset{n\to+\infty}{\longrightarrow} 0$ car on a supposé $d_n >> \log n$.
Si $m = m^*$, il est clair que $J_2(m) \underset{n\to+\infty}{\longrightarrow} |m^*|\,\sigma_*^2$ car $p_n(m^*) \underset{n\to+\infty}{\longrightarrow} 1$.
Par ailleurs, si $m \neq m^*$, on peut facilement majorer $J_2(m)$ en utilisant l'inégalité de Cauchy-Schwarz, puisque :

$$J_2 \leq \left[\mathbb{E}_{m^*}\left((\|P_{[X^{(m)}]}\varepsilon\|^2)^2\right)\right]^{1/2} \left[p_n(m)\right]^{1/2},$$

d'après la définition de $p_n(m)$. Or on montre facilement que

$$\mathbb{E}_{m^*}\left((\|P_{[X^{(m)}]}\varepsilon\|^2)^2\right) = \mathbb{E}_{m^*}\Big(\sum_{i,j,i',j'=1}^{n} \pi_{ij}\pi_{i'j'}\varepsilon_i\varepsilon_j\varepsilon_{i'}\varepsilon_{j'}\Big).$$

En dénombrant les cas où deux ou 4 indices sont égaux, on montre facilement que $\mathbb{E}_{m^*}\left((\|P_{[X^{(m)}]}\varepsilon\|^2)^2\right)$ est bornée. Comme pour $m \neq m^*$, $p_n(m) \underset{n\to+\infty}{\longrightarrow} 0$, on en

déduit que $R_n(m) \underset{n\to+\infty}{\longrightarrow} 0$ tandis que $R_n(m) \underset{n\to+\infty}{\longrightarrow} |m^*|\,\sigma_*^2$. D'où le résultat final. ■

Conséquence : Pour le critère *GIC* et sous les hypothèses précédentes, le modèle sélectionné converge vers le vrai modèle, et d'après la propriété montrée précédemment,

$$R_n \underset{n\to+\infty}{\longrightarrow} |m^*| \cdot \sigma_*^2,$$

qui est le risque quadratique que l'on aurait obtenu si l'on avait eu la connaissance a priori (par exemple si un "oracle" s'était prononcé) du vrai modèle.

Remarque : Les premiers résultats concernant la convergence des différents critères ont été obtenus par Nishi (1984) dans le cadre gaussien. Depuis, de nombreux travaux de recherche ont porté sur la sélection de modèle en régression linéaire. Ces travaux (citons par exemple ceux de Baraud, 2000 [7], Barron *et al.*, 1999 [14] ou Birgé et Massart, 2001 [10]) étendent les résultats à des modèles plus généraux, non forcément gaussiens..., mais s'appuient sur des résultats théoriques dépassant largement le niveau de cet ouvrage. Depuis 2005, de nouveaux travaux cherchent à définir un critère de type GIC (9.11) a l'aide d'une estimation de (c_n) à partir des données. On citera notamment les travaux de Lavielle (2005) et Arlot et Massart (2009). Enfin, de nouveaux critères fondés plutôt sur une distance $\mathbb{L}^1$ tels le LASSO ont été également développés depuis les années 2000, critères permettant notamment de traiter le cas où le nombre de variables à choisir peut être plus grand que celui des individus (voir par exemple Meinshausen et Bühlmann, 2006).

7 Un algorithme de sélection de modèles

Les critères de sélection de modèles présentés travaillent à partir de la famille de modèles $\mathcal{M}$. Lorsque, pour des raisons logiques (ascendances entre les variables par exemple) ou des raisons heuristiques (lois physico-chimiques par exemple), on peut assigner a priori une relation de priorité entre les différentes variables explicatives, on supposera alors que la famille $\mathcal{M}$ est une famille hiérarchique. Dans ce cas-là, lorsque l'on dispose de k variables explicatives, la famille $\mathcal{M}$ comprend donc k modèles possibles et le choix d'un modèle par un des critères précédents demande le calcul du critère pour chacun des k modèles, ce qui n'est pas coûteux en temps de calcul.

Mais le cas général est le plus souvent celui d'une famille de modèles $\mathcal{M}$ dite exhaustive dans laquelle aucune variable explicative n'a a priori plus d'importance qu'une autre. Il faudra donc pour sélectionner un modèle prédictif calculer 2^k valeurs du critère choisi (toujours dans le cas où l'on dispose de k variables explicatives). Dans

le cas où k est grand cela peut devenir rapidement prohibitif (par exemple, si $k = 20$, il faudra calculer environ 10^6 différentes valeurs du critère), d'autant que si l'on a sélectionné une variable pour un modèle à p variables, elle ne sera pas forcément sélectionnée lorsque l'on considérera un modèle à $p + 1$ variables (c'est là une des conséquences de la recherche d'un modèle prédictif et non explicatif).

Furnival et Wilson en 1974 (voir [29]) ont proposé un algorithme "accélérant" la procédure de sélection d'un modèle prédictif. D'une manière générale, cet algorithme très puissant permet de trouver la meilleure régression à p régresseurs parmi k. Lorsque pour p fixé, le nombre de modèles possibles est C_k^p (donc un nombre devenant rapidement très important lorsque k et p sont grands), l'algorithme de Furnival et Wilson explore un nombre polynomial de modèles. Son principe est de posséder un critère permettant de couper dans l'arbre de recherche des modèles possibles.

Cet algorithme devra donc être utilisé en complément des critères de sélection de modèle prédictif et de la manière suivante. L'algorithme commence par donner la meilleure régression (dans un sens à définir) à un régresseur, puis à deux régresseurs, puis à trois régresseurs, etc... Il reste à fixer p, pour cela on peut utiliser un des critères C_p de Mallows, AIC, AIC_c ou BIC. Remarquons que cet algorithme est mis en œuvre dans les commandes R et SAS ci-dessous.

8 Exemple traité par logiciels informatiques

Nous reprenons ici le jeu de donnée concernant la chenille processionnaire. Le but est de choisir parmi les différentes variables explicatives celles qui ont une influence réelle sur la variable $X11$ (ou $X12$), nombre moyen de nids de chenilles (ou son logarithme).

Logiciel SAS :

On suppose les données rangées dans le data "sasuser.process" et on lance le programme suivant :

```
proc reg data=sasuser.process;
model X11=X1-X10/selection =rsquare best=1 cp aic bic;
model X11 X12=X1 X2 X4 X5/tol vif r;
plot r.*p.;
run; quit;
```

Le programme est volontairement chargé en options pour illustrer les possibilités

de SAS. La première ligne n'appelle pas de commentaire, dans la seconde, l'option "`selection = rsquare best =1`" revient à faire tourner l'algorithme de Furnival et Wilson. Cette option qui permet le tri automatique de régresseur est la meilleure qui soit, et SAS supporte sans problèmes jusqu'à 15 régresseurs ce qui couvre de nombreuses applications. Elle est clairement préférable aux options `backward`, `forward`, `stepwise` et `maxsquare` qui ont le même but, mais avec une autre méthodologie. Il reste à choisir la taille optimale du modèle. Pour ce faire, nous avons demandé l'impression du Cp de Mallows (commande `cp`), ainsi que les critères d'Akaike AIC (commande `aic`) et de Akaike-Schwarz BIC (commande `bic`). La troisième ligne utilise le modèle retenu par la ligne précédente avec le critère du Cp. On a illustré diverses possibilités : celle de mettre plusieurs variables à expliquer, de demander des diagnostics de multicolinéarité et la valeurs des résidus. Le graphique qui suit est en basse résolution, c'est le graphique classique résidus (r.) contre valeur prédite(p.). La sortie (nous ne donnons ci-dessous la sortie que pour la variable $X11$, celle pour $X12$ a exactement le même format) est la suivante :

```
Number in
Model R-Square      C(p)         AIC         BIC    Variables in Model
1     0.3407     16.7050    -23.1304    -22.4287    X9
-------------------------------------------------------------------------------
2     0.4723      9.7860    -28.2514    -26.9085    X1 X9
-------------------------------------------------------------------------------
3     0.5798      4.4953    -33.5415    -30.5878    X1 X2 X9
-------------------------------------------------------------------------------
4     0.6303      3.0699    -35.6398    -31.0283    X1 X2 X4 X5
-------------------------------------------------------------------------------
5     0.6584      3.1630    -36.1714    -29.8843    X1 X2 X4 X5 X9
-------------------------------------------------------------------------------
6     0.6734      4.1450    -35.6096    -27.7754    X1 X2 X4 X5 X9 X10
-------------------------------------------------------------------------------
7     0.6847      5.3830    -34.7300    -25.2789    X1 X2 X3 X4 X5 X9 X10
-------------------------------------------------------------------------------
8     0.6892      7.0782    -33.1895    -22.3997    X1 X2 X3 X4 X5 X6 X9 X10
-------------------------------------------------------------------------------
9     0.6902      9.0046     31.3015    -19.3800    X1 X2 X3 X4 X5 X6 X8 X9 X10
-------------------------------------------------------------------------------
10    0.6903     11.0000    -29.3085    -16.3334    X1 X2 X3 X4 X5 X6 X7 X8 X9 X10
```

Cette partie est la réponse à la premiere ligne de commande et illustre le choix de modèles. L'option `selection=rsquare best=1` donne le meilleur modèle à 1 régresseur, puis celui à 2 régresseur, etc... Dans cet exemple, on peut remarquer que les ensembles

de variables choisies ne sont pas hiérarchiques : par exemple, on ne passe pas du modèle sélectionné à 3 régresseurs à celui en contenant 4 en rajoutant simplement une variable (on en a remplacé une autre...).

Le vrai problème, ensuite, est de choisir la taille du modèle. Dans cet exemple, les critères de Cp et BIC minimum sont cohérents et choisissent un modèle à 4 régresseurs. En revanche, le critère AIC semble privilégier un modèle avec 5 régresseurs. Comme on l'a déjà signalé au cours du chapitre, on préférera choisir le modèle avec le critère BIC que AIC, et on sélectionnera ainsi le modèle avec les 4 variables $X1$, $X2$, $X4$ et $X5$.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	13.10487	3.27622	11.51	<.0001
Error	27	7.68670	0.28469		
corrected Total	31	20.79157			

Root MSE	0.53357	R-Square	0.6303
Dependent Mean	0.81406	Adj R-Sq	0.5755
Coeff Var	65.54360		

Parameter Estimates

Var.	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Int.	1	6.60309	1.02423	6.45	<.0001	.	0
X1	1	-0.00281	0.00078216	-3.60	0.0013	0.8949	1.11733
X2	1	-0.04565	0.01346	-3.39	0.0022	0.9797	1.02067
X4	1	-0.75510	0.21591	-3.50	0.0016	0.1763	5.67193
X5	1	0.16847	0.05154	3.27	0.0029	0.1809	5.52696

Obs	Dep Var Y	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2
1	2.3700	1.6964	0.1625	0.6736	0.508	1.325	|**|
2	1.4700	1.2602	0.1799	0.2098	0.502	0.418	| |
3	1.1300	1.3642	0.2090	-0.2342	0.491	-0.477	| |
:	:	:	:	:	:	:	: :

Remarquons d'abord que le graphique de résidus pour $X11$ est pathologique (les résidus sont plutôt positifs pour les faibles et fortes valeurs prédites et négatifs pour

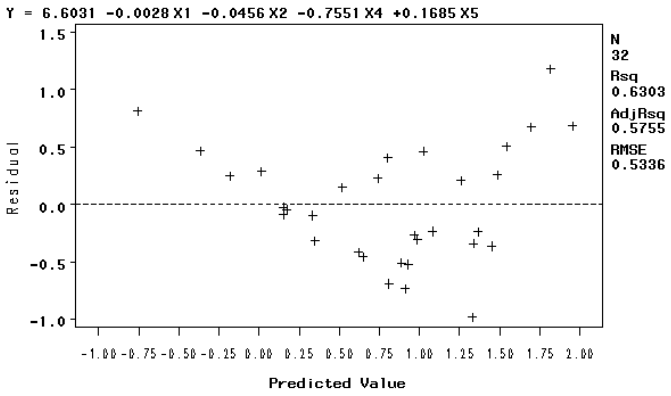


FIGURE 9.2 – Graphique SAS des résidus en fonction des valeurs prédites pour la régression $X111 \sim X1 + X2 + X4 + X5$ obtenue à partir de la sélection de modèle par critère Cp et BIC

les valeurs moyennes...). Que l'on se rassure, celui pour $X12$ ne l'est pas ! C'est pour cela que l'on a travaillé sur les logarithmes dans la première analyse.

Les valeurs des estimateurs montrent un effet positif (c'est-à-dire) une diminution) de l'altitude de la pente et de la hauteur des arbres. Cela peut éventuellement s'interpréter comme une difficulté d'accès à certaines placettes qui seraient protectrices. On constate que toutes les variables retenues par la procédure précédente sont significatives. Les indicateurs de colinéarité (on vérifie bien que TOL est l'inverse de VIF...) sont relativement raisonnables.

Les deux derniers tableaux sont dus à l'analyse de résidus demandé par l'option "/r". Le premier ne pose pas de problème d'interpretation. Le second est conçu pour la recherche de corrélations entre résidus consécutifs. Il comprend une représentation des résidus par un diagramme en bâtons, ainsi que la valeur du D de Cook : mesure d'influence de la mesure sur le paramètre estimé.

Logiciel R :

Reprenons les mêmes objectifs que précédemment :

```
library(leaps)                          library(car)
proc1=data.frame(proc[,1:10])           library(MASS)
```

```
r=leaps(proc1,proc$X11,nbest=10)
r$Cp; r$whi
plot(r$size-1,r$Cp)
t=(r$Cp==min(r$Cp))
colnames(proc1)[r$whi[t]]v
```

```
proc.lm=lm(proc$X11~.,proc1)
y.aic=stepAIC(proc.lm,k=2)
Anova(y.aic,type="III")
y.bic=stepAIC(proc.lm,k=log(32))
summary(y.bic)
par(mfrow=c(2,2))
plot(y.bic)
```

En premier lieu, notons la nécessité de faire appel aux deux librairies supplémentaires qui permettent d'utiliser les commandes `leaps` et `stepAIC`. Ensuite, on peut remarquer qu'il faut plusieurs commandes pour obtenir le modèle sélectionné par le critère du Cp de Mallows (on a cependant demandé en plus le tracé des valeurs du Cp en fonction du nombre de paramètres de tous les modèles d'analyse considérés). La détermination des modèles sélectionnés par les critères AIC et BIC se fait par la même commande `stepAIC`, mais avec un paramètre $k = log(n)$ pour le critère BIC, alors que par défaut, $k = 2$, ce qui correspond à la valeur de k pour le critère AIC. On peut d'ailleurs choisir un autre paramètre (comme par exemple, $k = log(log(n))$ qui peut être également intéressant, voir [50]). Voici des extraits des résultats des commandes concernant le Cp de Mallows :

```
> r$Cp; r$whi
 [1] 16.705043 17.471322 18.019757 19.091645 20.499496 25.164678 31.140784
 [8] 36.727516 38.121380 38.794452  9.785990 10.517001 11.333221 12.111766
[15] 12.557535 13.195754 13.532373 14.022935 14.842190 16.830370  4.495299
[22]  6.250653  8.287642  9.230832  9.619789  9.888255 10.991778 11.076856
 :       :         :         :         :         :         :         :

       1     2     3     4     5     6     7     8     9     A
1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
1  FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
:           :           :           :           :           :

> colnames(proc1)[r$whi[t]]
[1] "X1" "X2" "X4" "X5"
```

On demande d'abord les valeurs du Cp et le modèle associé pour les 10 modèles d'analyse ayant le plus faible Cp parmi tous les modèles possibles à nombre fixé de variables, en faisant croître progressivement ce nombre de variables (l'option $nbest = 10$, donnée également par défaut, peut être modifiée). Par exemple, la première valeur du Cp, soit $\simeq 16.705043$, correspond au modèle `FALSE FALSE FALSE FALSE FALSE FALSE FALSE`

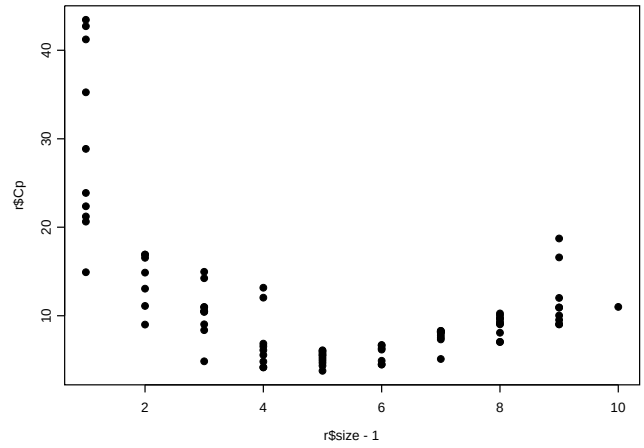


FIGURE 9.3 – Graphe des valeurs du critère Cp pour certains modèles en fonction du nombre de variables explicatives intervenant dans ces modèles (seuls les 10 "meilleurs" modèles sont représentés pour un nombre de variables fixé)

`FALSE TRUE FALSE`, c'est-à-dire au modèle $X11 \sim X9$: **on retrouve exactement les mêmes résultats qu'avec le logiciel SAS**. Le modèle finalement sélectionné par le critère du Cp de Mallows est $X11 \sim X1 + X2 + X4 + X5$; on pourra examiner également le graphe 8. Voici maintenant des extraits des résultats des commandes concernant les critères AIC et BIC :

```
> y.aic=stepAIC(proc.lm,k=2)
Start:  AIC= -29.31
 proc$X11 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq      RSS      AIC
- X7    1    0.0014   6.4402 -31.3015
- X8    1    0.0193   6.4582 -31.2125
- X6    1    0.0467   6.4855 -31.0774
- X9    1    0.1288   6.5677 -30.6745
- X3    1    0.1643   6.6031 -30.5022
<none>                6.4388 -29.3085
- X10   1    0.4446   6.8834 -29.1720
- X4    1    0.6163   7.0551 -28.3834
```

```
- X5    1    1.1383   7.5772 -26.0991
- X2    1    2.5198   8.9586 -20.7399
- X1    1    2.8463   9.2852 -19.5941

:           :           :           :
:           :           :           :

Step:  AIC= -36.17
 proc$X11 ~ X1 + X2 + X4 + X5 + X9

       Df Sum of Sq     RSS     AIC
<none>                7.102 -36.171
- X9    1     0.585   7.687 -35.640
- X4    1     1.477   8.579 -32.127
- X5    1     1.616   8.718 -31.609
- X2    1     2.535   9.637 -28.404
- X1    1     2.852   9.954 -27.367

> Anova(y.aic,type="III")
Anova Table (Type III tests)

Response: proc$X11
             Sum Sq Df F value    Pr(>F)
(Intercept) 11.1941  1 40.9810 8.808e-07 ***
X1           2.8523  1 10.4421  0.003332 **
X2           2.5351  1  9.2808  0.005256 **
X4           1.4766  1  5.4056  0.028146 *
X5           1.6164  1  5.9177  0.022177 *
X9           0.5847  1  2.1405  0.155443
Residuals    7.1020 26

> y.bic=stepAIC(proc.lm,k=log(32))
Start:  AIC= -13.19
 proc$X11 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq      RSS      AIC
- X7    1    0.0014   6.4402 -16.6442
- X8    1    0.0193   6.4582 -16.5551

:           :           :           :
:           :           :           :
```

```
Step:  AIC= -28.31
 proc$X11 ~ X1 + X2 + X4 + X5

       Df Sum of Sq      RSS      AIC
<none>                 7.6867 -28.3111
- X5    1    3.0422  10.7289 -21.1066
- X2    1    3.2755  10.9622 -20.4182
- X4    1    3.4821  11.1688 -19.8205
- X1    1    3.6820  11.3687 -19.2529
```

La commande `stepAIC` affiche donc la suite décroissante (au sens de l'inclusion) des modèles choisie à partir du critère et en partant du modèle global. En ce sens, cette procédure ne considère pas la famille exhaustive de modèles, mais une famille hiérarchique de modèles (le choix de cette famille est aléatoire et dépend des valeurs du critère). On remarque que l'itération s'arrête pour le critère AIC au modèle $X11 \sim X1 + X2 + X4 + X5 + X9$ pour lequel $AIC \simeq -36.17$ (on retrouve le même résultat qu'avec le logiciel SAS) modèle non validé par test de Fisher (la P-value de $X9$ est supérieure à 0.05).

En ce qui concerne la suite de modèles sélectionnés par le critère BIC, on peut remarquer que l'on aboutit au modèle $X11 \sim X1 + X2 + X4 + X5$ (modèle validé par test de Fisher) tout comme avec le logiciel SAS. En revanche, **la valeur elle-même du critère est différente : $BIC \simeq -28.31$ avec le logiciel R, alors que $BIC \simeq -31.03$ avec le logiciel SAS. On ne retrouve pas avec le logiciel SAS la valeur usuelle du coefficient BIC, puisque celui-ci est obtenu suivant une formule différente, ne faisant notamment pas appel à $\log n$, que l'on trouve par exemple dans l'article de Sawa (1978).**

Remarque : Les logiciels admettent parfois suivant les commandes différentes définitions pour les différents critères. Par exemple,

- **le critère Cp de Mallows** calculé à partir des commandes `proc reg` de SAS et `leaps` de R, vérifie :

$$Cp(m) = \frac{\tilde{\sigma}^2_{(m)}}{\tilde{\sigma}^2_{(m_k)}} + 2\,\frac{|m|}{n},$$

où $|m|$ la taille du modèle (nombre de variables), $\tilde{\sigma}^2_{(m)} = \dfrac{1}{n} \cdot SCR$ est l'estimateur du maximum de vraisemblance de σ^2 pour le modèle m et m_k est le modèle global contenant toutes les variables.

– avec les mêmes notations, **le critère** AIC calculé à partir des commandes `proc reg` de SAS et `stepAIC` de R, utilise la définition :

$$AIC(m) = n \cdot \tilde{\sigma}^2_{(m)} + 2 \cdot (|m| + 1).$$

– avec les mêmes notations, **le critère** BIC calculé dans la commande `stepAIC` de R (en précisant le paramètre $k = log(n)$) est tel que :

$$BIC(m) = n \cdot \tilde{\sigma}^2_{(m)} + \log(n) \cdot (|m| + 1).$$

– **les critères** AIC **et** BIC calculés par les commandes `AIC` pour le logiciel R (on rajoute l'option `k=2` pour le critère AIC et `k=log(n)` pour le critère BIC) sont issus des formules :

$$n \cdot \log(2\pi) - n + n \cdot \tilde{\sigma}^2_{(m)} + k \cdot (|m| + 2),$$

expression issue de la formule générale $-2\log(Vraisemblance)+k\cdot$nombre de param., en ayant supposé le modèle gaussien.

– enfin, comme nous l'avons déjà signalé, **le critère** BIC calculé à partir de la commande `proc reg` de SAS est celui proposé par Sawa (1978), soit :

$$BIC(m) = n \cdot \tilde{\sigma}^2_{(m)} + 2(|m| + 1)\frac{\tilde{\sigma}^2_{(m_k)}}{\tilde{\sigma}^2_{(m)}} - 2\left(\frac{\tilde{\sigma}^2_{(m_k)}}{\tilde{\sigma}^2_{(m)}}\right)^2.$$

Chapitre 10

Modèles mixtes

Dans ce chapitre, nous décrivons les modèles mixtes, qui sont des modèles à effets fixes et aléatoires. Ces modèles servent à décrire des expériences à plusieurs sources d'erreurs. Nous étudions les méthodes d'estimation et de test.

1 Modèles mixtes équirépétés

1.1 Un exemple

Reprenons l'exemple des insecticides déjà évoqué lors du chapitre 5. On désire comparer les aptitudes de 4 firmes à produire des insecticides efficaces. Chaque firme produit de nombreux insecticides, mais on échantillonne exactement deux produits numérotés (1 et 2) par firme. Pour étudier les insecticides, on utilise 24 boîtes contenant 400 moustiques chacune. Chaque produit est introduit dans 3 boîtes prises au hasard parmi les 24 et on compte le nombre de moustiques survivants au bout d'un temps déterminé. On considère que le nombre 400 est suffisamment grand pour que l'aspect binomial discret du problème puisse être oublié. Cela permet d'éviter d'utiliser un modèle linéaire généralisé à effets aléatoires. On reconnaît une structure dans laquelle l'effet `produit` est hiérarchisé au facteur `firme`. Comme les produits ont été échantillonnés, le facteur `produit` est maintenant aléatoire.

On veut répondre aux questions : existe-t-il une différence entre les firmes ? Quelle est la variabilité la plus importante : celle relative au choix du produit ou la variabilité résiduelle ?

Pour l'instant nous allons voir comment répondre à la seconde question. On note Y_{ijl} l'observation dans la l-ème boite du j-ème produit de la firme i, pour $i = 1, \cdots, I$; $j = 1, \cdots, J$ et $l = 1, \cdots, L$ (on pourra prendre par exemple par la suite $I = 4$, $J = 2$ et $L = 3$). On pose le modèle

$$Y_{ijl} = \mu + f_i + P_{ij} + \varepsilon_{ijl} \tag{10.1}$$

où P_{ij} désignent les effets aléatoires du produit hiérarchisé à firme. On suppose qu'ils sont tous centrés, gaussiens et indépendants de variance σ_P^2. On suppose également que les erreurs ε_{ijl} sont indépendantes des P_{ij} de variance σ^2 et vérifient les hypothèses classiques du modèle linéaire .

L'importance relative des effets va s'apprécier par l'estimation des différentes variances. Ici on est dans un cas équirépété, l'estimation est très simple. Il existe un estimateur qui est uniformément meilleur comme on peut le montrer par des techniques d'exhaustivité (on trouvera plus de précisions sur ce thème dans Coursol, 1980 [19]). On définit la somme des carrés résiduelle SCR et la somme des carrés associée au facteur `produit`, que l'on note SCF. Par utilisation de l'orthogonalité du modèle, il est facile de vérifier que

$$SCR = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{L}(Y_{ijl} - Y_{ij.})^2 \tag{10.2}$$

$$SCF = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{l=1}^{L}(Y_{ij.} - Y_{i..})^2 . \tag{10.3}$$

On laisse à titre d'exercice de montrer que

$$\mathbb{E}(SCR) = I.J(L-1)\sigma^2 \tag{10.4}$$

$$\mathbb{E}(SCF) = I(J-1)L\sigma_P^2 + I(J-1)\sigma^2 \tag{10.5}$$

(Indication : pour SCR c'est un résultat classique ; pour SCF il suffit de remarquer que cette somme est, à un facteur près, la somme des carrés du modèle basé sur les $Y_{ij.}$, qui est un modèle d'analyse de la variance à un facteur).

On rappelle ici la difficulté nouvelle que pose la présence d'un terme aléatoire (P_{ij}) supplémentaire dans le modèle. Pour proposer des estimateurs des variances inconnues σ_P^2 et σ^2, on va utiliser *la méthode des moments*, qui consiste à identifier les espérances aux valeurs effectivement observées. Plus précisément, ayant calculé les valeurs $\widehat{SCR}$ et $\widehat{SCF}$ à partir des données (donc comme indiqué en (10.2) et (10.3)), on détermine $\widehat{\sigma_P^2}$ et $\widehat{\sigma^2}$ en résolvant le système d'équations

$$\begin{aligned}\mathbb{E}(SCR) &= \widehat{SCR}\\ \mathbb{E}(SCF) &= \widehat{SCF},\end{aligned}$$

soit le système triangulaire inversible

$$\begin{aligned}\widehat{SCR} &= I.J(L-1)\widehat{\sigma}^2 \\ \widehat{SCF} &= I(J-1)L\widehat{\sigma}_P^2 + I(J-1)\widehat{\sigma}^2.\end{aligned}$$

Comme le cas équirépété reste un cas particulier, nous ne détaillerons pas, mais on montre (voir encore Coursol, 1980) que dans ce cas, les estimateurs ci-dessus sont optimaux parmi les estimateurs sans biais.

1.2 Fixe ou aléatoire ?

Nous allons nous livrer à un peu de considération méta-statistique. Dans beaucoup de modélisation, une des choses les plus difficiles à déterminer est le caractère fixe ou aléatoire d'un effet. En dehors des réponses "techniques" correspondant aux expériences randomisées et normalisées (voir le chapitre 11), voici la réponse tout à fait personnelle des auteurs.

On peut considérer que toute la statistique inférentielle a pour but d'apprécier la variabilité d'échantillonnage : est-ce que la différence entre les traitements (c'est la terminologie classique) que j'ai observée est due au hasard ou est-ce-qu'elle est due à un effet réel ? En d'autres termes, si on fait une répétition de l'expérience, de combien varieront les estimateurs ?

Cette dernière question prétend étendre le résultat de l'expérience à une population d'expériences. C'est la nature de cette population qui va déterminer le modèle :

- Si les niveaux d'un facteur ne changent pas de nature dans la répétition, alors on déclarera l'effet comme fixe ;

- Si les niveaux sont ré-échantillonnés dans une population, alors on déclarera l'effet comme aléatoire.

Illustrons ce "principe" sur l'exemple des insecticides.

Si chaque firme ne produit exactement que deux produits, alors ce sont ces mêmes produits qui seront utilisés dans une répétition de l'expérience et l'effet `produit` doit être déclaré comme un effet fixe.

Si chaque firme produit plus de deux produits, alors le choix des deux produits (choix qui est différent à chaque répétition) dans l'ensemble des produits est une nouvelle source de variabilité de l'expérience. L'effet `produit` doit être déclaré en aléatoire.

1.3 Modèles mixtes généraux

Première remarque : la méthode de la section 1 reste toujours applicable. Si on utilise les sommes de carrés de type III (voir chapitre 5) pour l'identification, la méthode porte alors le nom de *Méthode de Henderson III*. Le calcul des coefficients dans les expressions (10.4) et (10.5) est plus complexe, mais des expressions matricielles peuvent être établies par certaines techniques (voir dans la section suivante), expressions qui peuvent être calculées par ordinateur. Cette méthode est parfois très acceptable, surtout pour les dispositifs peu déséquilibrés, et dans tous les cas, elle fournit un bon point de départ pour les méthode itératives que nous allons développer un peu plus loin.

Au départ un modèle mixte est la donnée d'un vecteur Y qui peut s'écrire

$$Y = X \cdot \theta + Z_1 \cdot \beta_1 + \cdots + Z_{k-1} \cdot \beta_{k-1} + \varepsilon, \tag{10.6}$$

où θ est un vecteur de paramètres inconnus, les matrices $X, Z_1, \cdots, Z_k$ sont des matrices connues possédant n lignes et un nombre de colonnes compris entre 1 et n, et $\beta_1, \cdots, \beta_k$ sont des vecteurs aléatoires centrés, gaussiens, indépendants, et formés de coordonnées qui sont des variables aléatoires indépendantes et de même variance (donc de loi $\mathcal{N}(0, \gamma_i)$ pour $i = 1, \cdots, k$, où γ_i est la variance commune des coordonnées de β_i). Le modèle est mixte dans le sens où il regroupe des effets déterministes (ou fixes), le vecteur θ, et des effets aléatoires, les vecteurs $\beta_1, \cdots, \beta_k$.

De cette définition, on déduit facilement que

$$\begin{aligned} \mathbb{E}(Y) &= X \cdot \theta \\ \mathrm{Var}(Y) &= V_\gamma := \sum_{i=1}^{k} \gamma_i \cdot Z_i \cdot Z_i' = \sum_{i=1}^{k} \gamma_i \cdot V_i, \end{aligned}$$

où l'on pose $V_i = Z_i \cdot Z_i'$. C'est cette forme que l'on va poser comme définition :

Définition 10.1 (modèles mixtes) *Un vecteur aléatoire gaussien Y de taille n est dit suivre un modèle mixte statistique si*

$$\begin{aligned} \mathbb{E}(Y) &= X \cdot \theta \\ Var(Y) &= V_\gamma := \sum_{i=1}^{k} \gamma_i \cdot V_i, \end{aligned}$$

où $X, V_1, \cdots, V_k$ sont des matrices connues, θ est un vecteur de p paramètres inconnus variant dans $\mathbb{R}^p$, $\gamma := (\gamma_1, \cdots, \gamma_k)$ un vecteur de k paramètres inconnus variant dans l'ensemble $S := \{\gamma : V_\gamma > 0\}$, où la dernière notation veut dire que la matrice est définie positive.

Ce modèle est noté : $Y \sim MM(X, V_1, \cdots, V_k)$. Les $\gamma_1, \cdots, \gamma_k$ sont appelées composantes de la variance.

On peut donner maintenant une écriture équivalente à celle du modèle (10.6) précédent, soit :

$$Y = X \cdot \theta + \varepsilon \quad \text{avec} \quad \text{Var}\,(\varepsilon) = V_\gamma \tag{10.7}$$

1.4 Estimation des effets fixes

Pour simplifier, on supposera que la matrice X est de plein rang (on a vu que l'on pouvait se ramener à une telle hypothèse le plus souvent). Si γ est connu, le modèle mixte se ramène a un modèle linéaire tel que nous les avons traités précédemment, à la nuance près que le vecteur d'erreur ε admet maintenant une matrice de covariance qui ne se réduit pas forcément au cas $\sigma^2 \cdot Id$. Cependant, on peut montrer que l'estimateur optimal parmi les estimateurs linéaires sans biais est :

$$\widehat{\theta} = (X' \cdot V_\gamma^{-1} \cdot X)^{-1} X' \cdot V_\gamma^{-1} \cdot Y. \tag{10.8}$$

Cet estimateur est appelé "estimateur des moindres carrés généralisés. Il a été déjà étudié dans l'exercice 6 ***

Démonstration : La matrice V_γ étant supposée définie positive, on sait (résultat classique de la diagonalisation des matrices symétriques) qu'il existe une matrice T, symétrique, que l'on notera par la suite $V_\gamma^{-1/2}$, telle que

$$T \cdot T = V_\gamma^{-1} \quad \text{et} \quad T \cdot V_\gamma \cdot T = Id.$$

On pose donc $\widetilde{Y} = T \cdot Y = T \cdot X \cdot \theta + T \cdot \varepsilon = \widetilde{X} \cdot \theta + \widetilde{\varepsilon}$. Il est facile de voir que $\text{Var}\,(\widetilde{\varepsilon}) = Id$ et donc que $\widetilde{Y}$ suit un modèle linéaire à variance connue. Dans ce modèle l'estimateur optimal est

$$\widehat{\theta} = ((\widetilde{X})' \cdot \widetilde{X})^{-1} \cdot \widetilde{X}) \cdot \widetilde{Y}$$

donc en remplaçant dans cette expression, on montre bien que :

$$\widehat{\theta} = (X' \cdot V_\gamma^{-1} \cdot X)^{-1} \cdot X' \cdot V_\gamma^{-1} \cdot Y. \qquad \blacksquare$$

Notons que le calcul de l'estimateur d'expression (10.8) est simple numériquement même pour des modèles comprenant plusieurs centaines de paramètres.

De la même manière, si on considère une hypothèse linéaire sur le vecteur θ, le test de Fisher dans le modèle $\widetilde{Y} = \widetilde{X} \cdot \theta + \widetilde{\varepsilon}$ sera optimal. Quand γ ne sera plus connu mais

estimé (par exemple par $\widehat{\gamma}$), on estimera encore les effets fixes en utilisant les équations ci-dessus en remplaçant γ par $\widehat{\gamma}$. En conclusion, **le problème statistique essentiel dans un modèle mixte est l'estimation des composantes de la variance.**

1.5 Estimation par MIVQUE dans un modèle mixte

Dans cette section ainsi que dans la suivante, nous donnons de manière assez résumée plusieurs résultats sans démonstration.

Supposons que l'on désire estimer une composante γ_i de γ ou plus généralement une combinaison linéaire

$$C_h = h' \cdot \gamma$$

des composantes (où $h \in \mathbb{R}^k$ est donné). On cherche un estimateur $\widehat{C_h}$ de C_h qui vérifie les propriétés suivantes :

(i) l'estimateur est invariant par rapport aux effets fixes : $\widehat{C_h} = \widehat{C_h}(Y - X \cdot \theta),\ \forall \theta \in \mathbb{R}^p$;

(ii) l'estimateur est quadratique : $\widehat{C_h} = Y' \cdot B \cdot Y$ où B est une matrice (estimateur construit à partir de sommes de carrés) ;

(iii) l'estimateur est sans biais : $\mathbb{E}\left(\widehat{C_h}\right) = C_h$;

(iv) l'estimateur est de faible variance.

Pour déterminer un tel estimateur, on peut s'aider de la propostion suivante :

Proposition 10.1 *Vis-à-vis des propriétés souhaitées pour l'estimateur $\widehat{C_h}$,*

i. On peut toujours supposer la matrice B de (ii) symétrique et on le fera tout le temps ;

ii. L'estimateur est invariant si et seulement si $B.X = 0$ (B est maintenant symétrique) ;

iii. Si Y est gaussien et si $B \cdot X = 0$, $Var(Y' \cdot B \cdot Y) = Var(\varepsilon' \cdot B \cdot \varepsilon) = 2\,Tr(B \cdot V_\gamma \cdot B \cdot V_\gamma)$.

MIVQUE réel

La difficulté principale pour déterminer un tel estimateur vient du fait que la variance de $\widehat{C_h}$ dépend en général de la valeur de γ que l'on cherche à estimer. Il est

donc sans espoir (sauf dans les cas équirépétés) de chercher un estimateur optimal (au sens de variance minimale). On va plutôt se donner un **a priori**

$$\gamma \simeq \gamma_0,$$

et on va minimiser la variance gaussienne au voisinage de γ_0, ce qui revient donc à minimiser $\mathrm{Tr}(B \cdot V_{\gamma_0} \cdot B \cdot V_{\gamma_0})$. Ceci amène au théorème suivant :

Théorème 10.1 *Soit γ_0 un a priori. Un estimateur vérifiant (i), (ii), (iii) et de γ_0-variance minimale est appelé MIVQUE (Minimum Variance Quadratic Unbiased Estimator). Il est donné par les équations*

$$\widehat{C_h}(Y) = Y' \cdot Q'_{\gamma_0} \cdot V_{\gamma_0}^{-1} \sum_{i=i}^{k} \delta_i \cdot V_i \cdot V_{\gamma_0}^{-1} \cdot Q_{\gamma_0} \cdot Y$$

avec

- $Q_{\gamma_0} := Id - X \cdot (X' \cdot V_{\gamma_0}^{-1} \cdot X)^{-1} \cdot X' \cdot V_{\gamma_0}^{-1}$;
- δ *la solution de* $\delta' \cdot M_{\gamma_0} = h$, *où* M_{γ_0} *est la matrice* k, k *définie par*
- $M_{\gamma_0} = \{\, Tr(V_i \cdot V_{\gamma_0}^{-1} Q_{\gamma_0} \cdot V_j \cdot V_{\gamma_0}^{-1} Q_{\gamma_0})$ *pour* $1 \leq i, j \leq k.$ $\}$

En particulier, si la matrice M_{γ_0} est inversible, il existe un MIVQUE de toute composante.

MIVQUE vectoriel

Supposons M_{γ_0} inversible. Comme toute combinaison est estimable, on obtient un MIVQUE de chaque coordonnée γ_i. Pour ce faire, on cherche les k vecteurs δ^i tels que

$$M_{\gamma_0} \cdot \delta^i = f_i$$

où f_i est le i-ème vecteur de la base canonique (avec un 1 en i-ème position). Cela implique que

$$\{\delta^i_j, 1 \leq i, j \leq k\} = (M_{\gamma_0})^{-1}.$$

On définit ensuite le vecteur des sommes de carrés :

$$S = \{Y' \cdot Q'_{\gamma_0} V_{\gamma_0}^{-1} V_l V_{\gamma_0}^{-1} Q_{\gamma_0} Y,\ l = 1, \cdots, k\}.$$

On obtient alors l'estimateur :

$$\widehat{\gamma} = (M_{\gamma_0})^{-1} S.$$

1.6 Estimation par maximum de vraisemblance restreinte

On suppose que X est de plein rang, Y est un vecteur gaussien. La densité (par rapport à la mesure de Lebesgue) de X vaut

$$(2\pi)^{-n/2}|V_\gamma|^{-1/2}\exp\left(\frac{1}{2}(Y-X\cdot\theta)'\cdot V_\gamma^{-1}\cdot(Y-X\cdot\theta)\right).$$

On passe à $-2\log$ de cette expression, et on est conduit à maximiser la log-vraisemblance

$$L(\theta,\gamma) := \log(|V_\gamma|) + (Y-X\cdot\theta)'\cdot V_\gamma^{-1}.(Y-X\cdot\theta),$$

afin d'obtenir les estimateurs par maximum de vraisemblance des vecteurs de paramètres θ et γ. Pour expliciter ces estimateurs, on remarque que

$$\frac{\partial V_\gamma^{-1}}{\partial\gamma_i} = -V_\gamma^{-1}\frac{\partial V_\gamma}{\partial\gamma_i}V_\gamma^{-1}$$

donc

$$\frac{\partial log|V_\gamma|}{\partial\gamma_i} = \mathrm{Tr}(\frac{\partial V_\gamma'}{\partial\gamma_i}V_\gamma^{-1}).$$

On en déduit que

$$\frac{\partial L(\theta,\gamma)}{\partial\theta} = 0 \Leftrightarrow (X'\cdot V_\gamma^{-1}\cdot X)\theta = X'\cdot V_\gamma^{-1}\cdot Y.$$

Cette équation est connue sous le nom d'*équation de Gauss-Markov* . Elle permet de retrouver l'équation (10.8).

$$\frac{\partial L(\theta,\gamma)}{\partial\gamma_i} = 0 \Leftrightarrow \mathrm{Tr}(V_i\cdot V_\gamma^{-1}) = (Y-X\cdot\theta)'\cdot V_\gamma^{-1}\cdot V_i\cdot V_\gamma^{-1}\cdot(Y-X\cdot\theta).$$

Cependant, dans beaucoup de cas, les simulations montrent que le maximum de vraisemblance est biaisé en ce qui concerne les composantes de la variance. Comme la vraie difficulté, on l'a vu, réside dans l'estimation de ces composantes, on va, d'une certaine manière, "concentrer" la vraisemblance sur cette estimation.

Définition 10.2 *On appelle vraisemblance restreinte, la vraisemblance de $T'.Y$ où T est une matrice quelconque à n lignes et de rang maximal telle que $T'\cdot X = 0$. Comme $T'\cdot Y \sim \mathcal{N}(0, T'\cdot V_\gamma\cdot T)$, on pose*

$$L_R(\gamma) := log(|T'\cdot V_\gamma\cdot T|) + Y'\cdot T\cdot(T'\cdot V_\gamma\cdot T)^{-1}\cdot T'\cdot Y.$$

Cette vraisemblance restreinte ne concerne plus que γ (θ n'intervient pas dans son expression). Une fois choisie T telle que $T'\cdot X = 0$, on montre que cette matrice n'intervient plus dans l'estimation par maximum de vraisemblance restreinte :

Proposition 10.2 *La vraisemblance restreinte L_R ne dépend de la matrice T qu'à travers l'ajout d'une constante qui ne change rien à la maximisation de L_R. De plus,*

$$\frac{\partial L_R(\gamma)}{\partial \gamma_i} = 0 \Leftrightarrow Tr(V_i \cdot V_\gamma^{-1} \cdot Q_\gamma) = Y' \cdot Q_\gamma . V_\gamma^{-1} \cdot V_i \cdot V_\gamma^{-1} \cdot Q_\gamma \cdot Y$$

En conséquence, en comparant les équations, on constate qu'un point fixe du MIVQUE ($\widehat{\gamma} = \gamma_0$) est une solution des équations du maximum de vraisemblance restreinte. L'itération du MIVQUE est donc une façon (parmi d'autres) de résoudre ces équations. Pour plus de détail, on pourra consulter Azaïs, Bardin et Dhorne (1993).

1.7 Tests

Pour ce qui concerne les tests d'hypothèses sur les effets fixes, on réalise des tests de Fisher en supposant que les estimateurs des composantes de la variance sont en fait les valeurs exactes et en utilisant la méthode de la section 1.4. Comme on a négligé l'erreur d'estimation des composantes de la variance, ces tests ne sont qu'approximatifs, et souvent non-conservatifs. Certaines alternatives sont proposées dans la littérature comme le test du rapport de vraisemblance ou le test de Kenward-Roger. Voir par exemple la notice de SAS pour plus de détails. Malheureusement, les avantages respectifs de ces tests ne sont pas clairs.

Le plus souvent, les tests sur les effets aléatoires correspondent à la nullité d'une variance : on peut chercher ainsi à tester la nullité d'un effet "famille génétique", ou "sujet" par exemple. La première solution consiste à utiliser un test exact de Fisher. En effet, la nullité d'un effet aléatoire correspond strictement à l'absence d'effet, c'est-à-dire également à la nullité d'un effet déclaré en effet fixe. En résumé, pour tester la nullité d'une composante de la variance, on peut déclarer l'effet correspondant en fixe et utiliser le test de Fisher correspondant. Sauf dans le cas équilibré (voir Coursol 1980), ce test n'est plus optimal. Mais il est exact, dans le sens où son niveau réel est toujours égal au niveau nominal.

Une autre option est d'utiliser les tests classiques asymptotiques associés à la méthode du maximum de vraisemblance : le test du rapport de vraisemblance et le test de Wald. Pour les présenter, nous utilisons les notations traditionnelles où θ est le paramètre du modèle statistique. Le θ de cette partie est donc en fait égal au vecteur de paramètre (θ, γ) du modèle mixte précédent dans le cas de la vraisemblance classique, il est égal au seul γ dans le cas de la vraisemblance restreinte.

Test du rapport de vraisemblance et test de Wald

Dans une expérience qui comprend un grand nombre de répétitions indépendantes et sous des hypothèses de régularité (modèles de vraisemblance réguliers) qui sont

vérifiées dans notre cas, on sait que l'estimateur du maximum de vraisemblance est asymptotiquement sans biais, normal et de variance donnée par l'inverse de l'information de Fisher (voir par exemple, Dacunha-Castelle et Duflo, 1990). L'information de Fisher $I(\theta)$ est donnée par

$$(I(\theta))_{ij} = -\mathbb{E}\left(\frac{\partial^2 \log V(\theta)}{\partial\theta_i \partial\theta_j}\right)$$

où V est la vraisemblance, θ le paramètre du modèle. Comme, $I(\theta)$ tend vers l'infini avec le nombre de répétitions, la phrase "asymptotiquement sans biais, normal et de variance donnée par l'inverse de l'information de Fisher" doit se comprendre comme

$$\big(I(\theta)\big)^{1/2}\big(\widehat{\theta} - \theta\big) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, Id),$$

quand le nombre de répétitions n tend vers l'infini.

Une telle propriété reste vraie pour la vraisemblance restreinte et dans des cadres plus généraux, mais sous des conditions très techniques (voir par exemple Jiang, 1996). Ces conditions sont en particulier que les composantes de la variance correspondent à des effets aléatoires $\beta_1, \cdots, \beta_k$ (au sens de notre première définition d'un modèle mixte) et que ces composantes soient toutes positives. Dans ces conditions, il est possible de construire *un test de Wald* de l'hypothèse

$$\theta_i = \theta_{i,0},$$

dont la région de rejet est

$$\left|\big(I(\theta)^{-1}_{ii}\big)^{-1/2}\big(\widehat{\theta_i} - \theta_{i,0}\big)\right| > Z_\alpha,$$

où Z_α est le quantile d'ordre $1-\alpha$ pour la valeur absolue d'une loi normale standard.

Compte tenu des hypothèses de l'article Jiang (1996), ce test n'est généralement pas valide pour tester la nullité d'une composante. En pratique, on peut vérifier que si on l'applique tout de même et ce surtout pour des petits échantillons, ce test est peu puissant et conservatif (le niveau réel est nettement plus important que le niveau nominal).

Des calculs élémentaires, quoique longs, montrent que dans le cas de la vraisemblance restreinte, on peut obtenir le même test, mais avec pour information de Fisher :

$$I_\gamma = \left(\frac{1}{2}\mathrm{Tr}(V_i \cdot V_\gamma^{-1} \cdot Q_\gamma \cdot V_j \cdot V_\gamma^{-1} \cdot Q_\gamma)\right)_{1\le i,j\le k}.$$

Un autre test possible qui est en général plus puissant quoique parfois non-conservatif, est le *test du rapport de vraisemblance* . Si L_g est la log-vraisemblance prise au maximum et L_p est la log-vraisemblance ou maximum sous l'hypothèse nulle, on montre

qu'avec les mêmes conditions que précédemment

$$L_g - L_p \xrightarrow[n\to+\infty]{\mathcal{L}} \frac{1}{2}\chi^2(k'),$$

où k' est la différence de dimension paramétrique entre les deux modèles : hypothèse générale, hypothèse nulle.

Ces deux tests, Wald et rapport de vraisemblance, peuvent être indifféremment utilisés pour le maximum de vraisemblance comme pour le maximum de vraisemblance restreinte. Cependant, pour de petits échantillons (notion qu'il est difficile de préciser), certains écarts entre le niveau réel et le niveau nominal font que le test de Fisher semble être un choix moins risqué.

2 Analyse de la variance multivariable

Dans cette section, nous donnons dans un premier temps une présentation classique de l'analyse de la variance multivariable, dans un second temps, nous montrons comment un tel modèle peut être vu - et donc analysé - comme un modèle mixte.

L'analyse de la variance, la régression et le modèle linéaire en général, ont pour but d'expliquer **une** variable quantitative par un certain nombre de variables explicatives quantitatives ou qualitatives. Dans de nombreux cas l'observation est multiple : on peut par exemple mesurer la longueur et la largeur d'une feuille ; le rendement d'une réaction chimique en plusieurs instants, etc... Une telle situation correspond souvent à la répétition d'une même mesure dans le temps, ce que l'on dénomme *données longitudinales* ou bien *plans à mesures répétées* .

L'analyse de la variance multivariable est l'une des méthodes les plus simples pour modéliser une telle situation : on va maintenant supposer que l'observation est vectorielle et qu'elle se situe dans $\mathbb{R}^L$, où $L \in \mathbb{N}^*$. Une analyse variable par variable est toujours possible. Elle donne des estimations qui sont parfaitement exploitables. Pour ce qui est des tests, la situation est plus complexe : chaque analyse univariable donne un test différent pour une hypothèse telle que, par exemple, celle de la nullité de l'effet d'un facteur. Ces tests sont peu puissants car ils n'utilisent que l'information d'une seule variable et ils peuvent même être contradictoires. Il est donc très souhaitable de pouvoir regrouper les tests sur les différentes variables.

Exemple : Pour apprécier et comparer le degré de pollution de rivières, on dose différents métaux lourds dans les viscères de truites. On dispose par exemple de 6

variables observées, que l'on peut noter $(Y^{(1)}, \cdots, Y^{(6)})$:

Plomb dans le foie	Plomb dans l'intestin	Plomb dans le colon
Mercure dans le foie	Mercure dans l'intestin	Mercure dans le colon.

Supposons que l'on observe plusieurs truites par rivières et notons Y_{ij} l'observation sur la truite j de la rivière i. La situation ressemble tout à fait à celle d'une analyse de la variance à un facteur à ceci près que l'observation se trouve dans $\mathbb{R}^6$. On pose ainsi :

$$Y_{ij} = t_i + \varepsilon_{ij} \text{ avec } t_i \in \mathbb{R}^6 \text{ et } \varepsilon_{ij} \sim \mathcal{N}(0, \Sigma) \text{ pour } i = 1, \cdots, I \text{ et } j = 1, \cdots, J, \quad (10.9)$$

où Σ est une matrice carrée d'ordre 6 définie positive. Comme nous l'avons déjà souligné, du fait de la linéarité du modèle, l'estimation peut parfaitement se faire variable par variable. En revanche, le test de la "nullité de l'effet rivière" (par exemple) aura 6 réponses en fonction de la variable utilisée. Pour obtenir un test unique, nous allons paraphraser les formules de l'analyse de la variance univariable en essayant de les généraliser au cas vectoriel.

On peut toujours estimer t_i par $Y_{i.}$ qui est le vecteur des moyennes. On définit ainsi le vecteur des résidus associé à l'unité i, j.

$$\widehat{\varepsilon}_{ij} = Y_{ij} - \widehat{t_i} = Y_{ij} - Y_{i.}. \quad (10.10)$$

Comme il s'agit toujours d'un vecteur de taille 6 (taille que l'on fixera maintenant à L pour généraliser), la somme des carrés (notée SCR dans les chapitres précédents) est remplacée par une matrice carrée d'ordre (L, L) contenant les différentes sommes des carrés et produits . Ainsi, en notant $SCR^{k,l}$ le terme de la k-ème ligne et l-ème colonne de cette matrice, on a

$$SCR^{(k,l)} = \sum_{i=1}^{I} \sum_{j=1}^{J} \widehat{\varepsilon}_{ij}^{(k)} \times \widehat{\varepsilon}_{ij}^{(l)} \ , \text{ pour } 1 \leq k, l \leq L \quad (10.11)$$

($\widehat{\varepsilon}_{ij}^{(k)}$ désigne la k-ème coordonnée du vecteur $\widehat{\varepsilon}_{ij}$). Dans le cas univariable, la variable aléatoire SCR suivait une loi $\sigma^2 \chi^2(n - I)$ ($n = I \times J$ est le nombre d'observations). Dans le même état d'esprit, on définit

Définition 10.3 (Loi de Whishart de paramètres d et Σ) *Soit d vecteurs aléatoires indépendants $X_1, \cdots, X_d$ de loi $\mathcal{N}(0, \Sigma)$. On définit la matrice M des sommes de carrés et produits par*

$$M = \sum_{i=1}^{d} X_i \cdot (X_i)'.$$

Cette matrice est composée de sommes de carrés et produits qui suivent des lois $\sigma^2 \cdot \chi^2(d)$ sur la diagonale et de sommes de produits à l'extérieur de la diagonale. Par définition elle suit la loi de Whishart de paramètres d et Σ.

Ainsi, en utilisant la normalité des erreurs, l'écriture sous forme de projection des $\widehat{\varepsilon}_{ij}$ et le Théorème de Cochran, on obtient :

Propriété 10.1 *Sous les hypothèses du modèle (10.6), la matrice aléatoire SCR est telle que :*

$$SCR \text{ suit une loi de } Whishart(N-I, \Sigma).$$

On définit également la matrice aléatoire dite somme de carrés et produits associée au facteur :

$$SCF := \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{i.} - Y_{..}).(Y_{i.} - Y_{..})'.$$

Sous l'hypothèse H_0 d'absence d'effet du facteur (rivière dans notre exemple), on montre que

$$SCF \text{ suit une loi de Whishart de paramètres}(I-1, \Sigma).$$

Il reste à construire le test de l'absence d'effet d'un facteur pour le vecteur d'observation. Rappelons qu'en analyse univariable, ce test est fondé sur le rapport

$$\widehat{F} := \frac{SCF/(I-1)}{SCR/(N-I)}$$

qui suit la loi de Fisher $F_{(I-1),(N-I)}$.
En multivariable, d'un certain coté, la situation est analogue puisque $SCR/(N-I)$ est un estimateur de Σ et d'un autre coté, la situation est différente, dans le sens qu'il n'est pas possible de faire le quotient de deux matrices.

On pose alors $W := SCR/(N-I)$ (W comme within) matrice des variations intra-facteur et $B := SCF/(I-1)$ (B comme between) matrice de variations inter-facteur. Les tests en multivariable sont multiples et fondés sur les statistiques suivantes :

- le λ de Wilks : $\frac{|W|}{|B+W|}$ qui donne le test du rapport de vraisemblance ;
- la trace de Pillai : $Tr(B \cdot (B+W)^{-1})$;
- la trace de Hotelling-Lawley : $Tr(B \cdot W^{-1})$;

- la plus grande valeur propre (PGVP) de Roy : PGVP de $(W^{-1}B)$.

Fort heureusement, dans la plupart des cas, ces tests donnent des résultats cohérents. Les lois connues et complexes sous H_0 de ces statistiques ont de très bonnes approximations par des lois de Fisher qui sont utilisées le plus souvent par les logiciels pour calculer les niveaux de signification.

2.1 Un exemple de modèle multivariable se ramenant à un modèle mixte

On reprend ici l'exemple de la pollution des rivières cité dans le texte. Voici tout d'abord des commandes SAS permettant une analyse de la variance multivariable :

```
proc glm data=sasuser.pollution;
class riviere;
model Hgf Hgi Hgc Pbf Pbi Pbc=riviere;
means riviere/Tukey;
manova h=riviere; run;
```

Nous allons montrer maintenant que le modèle (10.9) est en fait un modèle mixte. Nous allons également donner les moyens de l'analyser avec R, SAS ou Splus. Nous regroupons toute les observations en un vecteur $(Y_{ij\ell})_{ij\ell}$ où $Y_{ij\ell}$ est l'observation de la variable ℓ sur le j-ème individu du groupe i. Il est clair que l'espérance est linéaire :

$$\mathbb{E}\left(Y_{ij\ell}\right) = t_{i\ell}$$

par ailleurs Σ peut être paramétrée linéairement par 21 paramètres : les 6 variances et les 15 covariances. Dans le cas général d'une observation dans $\mathbb{R}^L$, il faut $L(L+1)/2$ paramètres. Il s'ensuit que la matrice de variance-covariance de Y numérotée en ordre lexicographique est une matrice bloc diagonale avec $I \cdot J$ blocs égaux à Σ. Elle peut donc encore être paramétrée par $L(L+1)/2$ paramètres (21 dans l'exemple). Nous obtenons encore un modèle mixte.

Voici (par exemple) la solution SAS. Au contraire de la solution précédente, les données doivent être organisées de la façon suivante : les différentes réponses doivent être regroupées en une seule variable que nous noterons `reponse`. On doit parallèlement créer une variable `variable` qui indique de quelle réponse il s'agit. Par exemple, si le début du fichier était dans sa forme initiale :

```
riviere repetition Hgf Hgi Hgc Pbf Pbi Pbc
      1          1  12  25   4   6  22  47
```

Il doit devenir dans la nouvelle forme :

riviere	repetition	variable	reponse
1	1	1	12
1	1	2	25
1	1	3	4
1	1	4	6
1	1	5	22
1	1	6	47

Nous supposerons que les données sous cette forme ont été sauvées dans le fichier `sasuser.pollution2`. L'analyse correspondante est :

```
Proc mixed data=sasuser.pollution2;
Class riviere variable;
model reponse=riviere*variable;
repeated variable subject=riviere*repetition/US;
run; quit;
```

La commande `repeated` suppose l'indépendance d'un sujet (`subject`) à l'autre. Il reste à modéliser la dépendance intra-sujet, c'est-à-dire, dans notre cas, la dépendance entre les différentes variables de départ. L'option `US` comme "Unstructured" indique que la matrice de variance-covariance est non structurée, ce qui implique comme seule contrainte que la matrice doit être symétrique et définie positive. Cela correspond bien au modèle (10.9) de l'analyse de la variance multivariable.

En fait la présentation sous forme de modèle mixte de l'analyse de la variance multivariée permet une plus grande souplesse : on peut spécifier (même si ce n'est pas trivial) certains sous-modèles du modèle général, mais en revanche on n'a pas accès aux différents tests (Wilks, Roy, etc...).

3 Exemples traités par logiciels informatiques

3.1 Modèle mixte

Nous revenons maintenant sur l'exemple des insectes du chapitre 5, mais nous supposons maintenant que chaque firme conçoit plus de deux produits. Ainsi que cela

a déjà été formulé dans l'introduction, les produits présents dans l'expérience ont été échantillonnés, ce qui conduit à déclarer l'effet `produit` comme aléatoire (voir Milliken et Johnson [47]).

Logiciel Splus :

La commande `menuLme` synthétise les instructions nécessaires à l'étude des modèles mixtes. Dans la déclaration du modèle, on doit d'abord indiquer les effets fixes, puis les effets aléatoires dans l'option `random` (l'écriture de ces effets se fait par groupement de variables, comme suit). Comme nous l'avons déjà évoqué plusieurs fois, **avec Splus, il faut lancer deux commandes `menuLme`, la première permettant d'effectuer correctement l'analyse de la variance, la suivante permettant l'estimation des différents coefficients** :

```
attach(insect)
produit<-as.factor(produit)
firme<-as.factor(firme)
menuLme(nb~firme,random=~1|produit/firme,print.anova.p=T)
menuLme(nb~firme-1,random=~1|produit/firme)
```

D'où les résultats (extraits) :

```
> menuLme(nb~firme,random=~1|produit/firme,print.anova.p=T)

    *** Linear Mixed Effects Model ***

Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  168.1444 175.1145 -77.07221

Random effects:
 Formula:  ~ 1 | produit
        (Intercept)
StdDev:    5.698435

 Formula:  ~ 1 | firme %in% produit
        (Intercept) Residual
StdDev:   0.6666831  7.72442
```

```
Analysis of Variance Table
            numDF denDF  F-value p-value
(Intercept)     1    16 644.3796  <.0001
      firme     3     3 109.1347  0.0015

> menuLme(nb~firme-1,random=~1|produit/firme)

    *** Linear Mixed Effects Model ***

Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  161.7883 168.7584 -73.89415

Random effects:
 Formula:  ~ 1 | produit
        (Intercept)
StdDev:    5.698519

 Formula:  ~ 1 | firme %in% produit
        (Intercept) Residual
StdDev:   0.6700751 7.724273

Fixed effects: nb ~ firme - 1
          Value Std.Error DF  t-value p-value
firmea 134.6667  5.138592  3 26.20692  0.0001
firmeb 141.1667  5.138592  3 27.47186  0.0001
firmec  91.8333  5.138592  3 17.87130  0.0004
firmed  72.3333  5.138592  3 14.07649  0.0008
 Correlation:
       firmea firmeb firmec
firmeb 0.615
firmec 0.615  0.615
firmed 0.615  0.615  0.615

Standardized Within-Group Residuals:
      Min         Q1        Med       Q3      Max
 -1.66525 -0.5454547 -0.1989568 0.419019 1.748821
```

Logiciel R :

Pour le logiciel R, on doit faire appel au "package" `nlme` pour travailler avec des effets aléatoires. Les commandes permettant d'obtenir les mêmes résultats qu'avec Splus (avec les mêmes inconvénients liés aux deux différentes analyses de la variance) sont :

```
library(nlme)
attach(insect)
produit=as.factor(produit)
firme=as.factor(firme)
insect.mix1=lme(nb~firme,random=~1|produit/firme)
anova(insect.mix1)
insect.mix2=lme(nb~firme-1,random=~1|produit/firme)
insect.mix2$coef
```

Logiciel SAS

La procédure de modèle mixte de SAS est `proc mixed`. En dehors de l'utilisation de la logique des plans à mesures répétées (voir section suivante) par la directive `repeated`, la syntaxe est très proche de `proc glm`. La principale différence réside dans le fait que :

- la ligne `model` déclare les effets fixes,
- la ligne `random` déclare les effets aléatoires.

Voici les commandes permettant d'introduire le modèle sur les insectes et de le traiter :

```
proc mixed data=sasuser.insect;
class firme produit;
model nb=firme;
random produit*firme;
lsmeans firme/tdiff;run; quit;
```

On obtient alors :

```
                    Iteration History

Iteration    Evaluations    -2 Res Log Like       Criterion
        0              1       151.41874428
```

1	1	149.60587025	0.00000000

Covariance Parameter Estimates

Cov Parm	Estimate
firme*produit	32.9167
Residual	59.6667

Fit Statistics

-2 Res Log Likelihood	149.6
AIC (smaller is better)	153.6
BIC (smaller is better)	153.8

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
firme	3	4	42.02	0.0018

Least Squares Means

Effect	firme	Estimate	Standard Error	DF	t Value	Pr > |t|
firme	a	134.67	5.1384	4	26.21	<.0001
firme	b	141.17	5.1384	4	27.47	<.0001
firme	c	91.8333	5.1384	4	17.87	<.0001
firme	d	72.3333	5.1384	4	14.08	0.0001

Les tests T ci-dessus sont sans intérêt. Les bons tests sont ci-dessous :

Differences of Least Squares Means

Effect	firme	firme	Estimate	Standard Error	DF	t Value	Pr > |t|
firme	a	b	-6.5000	7.2667	4	-0.89	0.4216
firme	a	c	42.8333	7.2667	4	5.89	0.0041
firme	a	d	62.3333	7.2667	4	8.58	0.0010
firme	b	c	49.3333	7.2667	4	6.79	0.0025
firme	b	d	68.8333	7.2667	4	9.47	0.0007

```
firme     c        d          19.5000      7.2667       4       2.68      0.0550
```

Commentaires : la convergence est immédiate du fait de l'équilibre du jeu de données. Il est difficile de faire un test correct de l'effet hiérarchisé de `produit`. Comme précisé, le meilleur choix reste le test de Fisher dans le modèle à effet fixe qui a été fait dans le chapitre 5.

Conclusion générale : L'effet `firme` est clairement significatif et on peut faire une comparaison des moyennes par la méthode de Bonferroni en utilisant la partie intitulée "Differences of Least Squares Means" (logiciel SAS uniquement, car nous n'avons pas trouvé d'équivalent pour modèle mixte avec Splus ou R). Il est à noter que les tests de Student des coefficients (appelée partie "Least Squares Means" en SAS), qui sont tous très significatifs, testent des hypothèses sans grand intérêt.

3.2 Analyse de la variance multivariable

Pour illustrer cette partie du cours, nous proposons un exemple extrait du livre de Krzanowski [38] (1998, p. 381), et proposé également dans l'aide du logiciel R. Il porte sur la production de films plastiques.

Logiciel R :

Voici la suite de commandes, avec notamment l'écriture des données :

```
tear=c(6.5, 6.2, 5.8, 6.5, 6.5, 6.9, 7.2, 6.9, 6.1, 6.3,
6.7, 6.6, 7.2, 7.1, 6.8, 7.1, 7.0, 7.2, 7.5, 7.6)
gloss=c(9.5, 9.9, 9.6, 9.6, 9.2, 9.1, 10.0, 9.9, 9.5, 9.4,
9.1, 9.3, 8.3, 8.4, 8.5, 9.2, 8.8, 9.7, 10.1, 9.2)
opacity <- c(4.4, 6.4, 3.0, 4.1, 0.8, 5.7, 2.0, 3.9, 1.9, 5.7,
2.8, 4.1, 3.8, 1.6, 3.4, 8.4, 5.2, 6.9, 2.7, 1.9)
Y=cbind(tear,gloss,opacity)
rate=factor(gl(2,10),labels=c("Low","High"))
additive=factor(gl(2,5,len=20),labels=c("Low","High"))
fit=manova(Y~rate*additive)
summary.aov(fit)
summary(fit,test="W")
summary(fit,test="P")
summary(fit,test="H")
summary(fit,test="R")
```

D'où les résultats (extraits) :

```
>summary.aov(fit)
Response tear :
              Df  Sum Sq Mean Sq F value   Pr(>F)
rate           1 1.74050 1.74050 15.7868 0.001092 **
additive       1 0.76050 0.76050  6.8980 0.018330 *
rate:additive  1 0.00050 0.00050  0.0045 0.947143
Residuals     16 1.76400 0.11025

 Response gloss :
              Df  Sum Sq Mean Sq F value  Pr(>F)
rate           1 1.30050 1.30050  7.9178 0.01248 *
additive       1 0.61250 0.61250  3.7291 0.07139 .
rate:additive  1 0.54450 0.54450  3.3151 0.08740 .
Residuals     16 2.62800 0.16425

Response opacity :
              Df Sum Sq Mean Sq F value Pr(>F)
rate           1  0.421   0.421  0.1036 0.7517
additive       1  4.901   4.901  1.2077 0.2881
rate:additive  1  3.961   3.961  0.9760 0.3379
Residuals     16 64.924   4.058
```

Les trois tables d'analyse de la variance univariées ci-dessus ne donnent pas des résultat identiques sur le plan qualitatif. Les test multivariables permettent une réponse sur les trois variables. Dans cet exemple, les tests W (Wilks), P (Pillai), h (Hotelling-Lawley) et R (Roy) donnent strictement le même résultat. Nous donnons seulement ce dernier :

```
> summary(fit,test="R")
              Df    Roy approx F num Df den Df   Pr(>F)
rate           1 1.6188   7.5543      3     14 0.003034 **
additive       1 0.9119   4.2556      3     14 0.024745 *
rate:additive  1 0.2868   1.3385      3     14 0.301782
Residuals     16
```

Logiciel Splus

Sur ce même exemple, les commandes permettant un traitement comparable sont

très similaires avec le logiciel Splus (si ce n'est que la commande `menuManova` permet également le tracé d'un grand nombre de graphes et le calcul de toute sorte d'estimations) :

```
attach(krza)
rate<-as.factor(rate)
additive<-as.factor(additive)
fit<-menuManova(Y~rate*additive,coef.p=T)
summary(fit,test="w")
summary(fit,test="p")
summary(fit,test="r")
summary(fit,test="h")
```

Logiciel SAS

Reprenons le même exemple des films plastiques traité cette fois-ci avec SAS :

```
proc glm data=sasuser.krza;
class rate additive;
model tear gloss opacity= rate additive rate*additive;
means rate additive rate*additive;
Manova h= rate additive rate*additive;
run; quit;
```

Les analyses univariées sont identiques à celle obtenues par R. Nous ne les donnons pas. Voici les moyennes calculée par la directive `means` (extrait violent) :

Level of rate	Level of additive	tear Mean	-gloss Mean	opacity Mean
High	High	7.28	9.40	5.020
High	Low	6.88	8.72	3.140
Low	High	6.68	9.58	3.840
Low	Low	6.30	9.56	3.740

La présentation des tests multivariés par SAS est différente de celle de R : pour un effet donné, par exemple l'interaction `rate*additive`, (par concision nous ne donnons pas les autre sortie), la directive `manova` donne dans un tableau la liste de différents tests. Le lecteur pourra vérifier leur redoutable cohérence.

Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1 tear	gloss	opacity
0.28682614	100.00	0.13635945	0.53757970	0.06825336
0.00000000	0.00	0.74900685	-0.01690768	-0.00214667
0.00000000	0.00	0.19526817	-0.30318939	0.11022451

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall rate*additive Effect
H = Type III SSCP Matrix for rate*additive
E = Error SSCP Matrix

S=1 M=0.5 N=6

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.77710576	1.34	3	14	0.3018
Pillai's Trace	0.22289424	1.34	3	14	0.3018
Hotelling-Lawley Trace	0.28682614	1.34	3	14	0.3018
Roy's Greatest Root	0.28682614	1.34	3	14	0.3018

Commentaires : Nous arrivons donc à la conclusion suivante : l'interaction n'est pas significative, les effets principaux le sont. Le modèle retenu est donc un modèles additif. Tout ça pour ça !

Chapitre 11

Présentation des plans d'expériences classiques

Dans ce chapitre, nous donnons d'abord un véritable plaidoyer pour la randomisation. Nous présentons ensuite les plans d'expériences randomisées classiques.

1 Introduction

Dans cette partie, nous allons présenter l'idée forte suivante : la statistique n'a pas comme seul objet de "traiter des données" mais également de préparer leur recueil pour améliorer leur "qualité". D'importants gains sont possibles lors de cette étape. Les méthodes que nous allons présenter s'appliquent à des expériences planifiées dans lesquelles les principales variables explicatives sont contrôlées, plutôt qu'à des situations où les données sont recueillies "comme elles viennent", ce qui est le cas, par exemple, dans les enquêtes.

Les deux buts de la planification sont :

1. De permettre une interprétation claire en évitant les confusions ;

2. De maximiser la précision de l'expérience.

Pour illustrer le point 1., prenons l'exemple de la scolarisation en maternelle. Des études incontestables ont montré que les enfants scolarisés en maternelle ont de meilleurs résultats dans la suite de leur scolarité que les enfants qui ne rejoignent

l'école qu'au primaire. Doit-on en déduire qu'il faut rendre la scolarisation en maternelle obligatoire pour lutter contre l'échec scolaire ? Une réponse directe par l'affirmative n'est pas possible. En effet, deux interprétations sont possibles : (a) c'est effectivement la scolarité en maternelle qui améliore les résultats ; (b) dans la France actuelle, les élèves qui ne vont pas en maternelle sont une minorité qui correspond à des groupes sociaux bien particuliers, ce qui peut expliquer la différence de réussite scolaire par l'origine sociale.
Dans cet exemple, on pourra affiner l'analyse en contrôlant toute les variables indiquant la situation sociale, avec toujours le risque d'en oublier une. Mais il est clair que sans cette information complémentaire, les données de départ sont sans valeur pour répondre à la question. Une solution de type planification à ce problème serait de définir un groupe d'enfants test et de les répartir "au hasard" en deux groupes, l'un qui serait scolarisé en maternelle, l'autre non. Bien sûr, pour des raisons éthiques, ceci est difficilement réalisable ; c'est ce que nous entendions comme la différence entre une expérience de "labo" et l'utilisation de données recueillies.

Pour illustrer le point 2., considérons le problème de la pesée de deux objets A et B avec une balance sans biais qui donne chaque résultat avec une erreur centrée indépendante et de variance σ^2. On suppose que la balance est capable également de peser les deux objets ensembles.

1. On pèse A et B séparément ; le coût de l'expérience est alors celui de deux pesées et la précision (variance de l'erreur) obtenue sur les pesées de A et de B est σ^2.

2. On pèse $A+B$ et $A-B$ (on suppose que A est plus lourd que B et que, grace à une balance de Roberval ou l'utilisation d'une double pesée, on peut peser $A-B$) ; le coût de l'expérience est encore de deux pesées et sa précision est $\sigma^2/2$, car les poids de $A = \frac{1}{2}[(A+B)+(A-B)]$ et $B = \frac{1}{2}[(A+B)-(A-B)]$ sont maintenant obtenus comme moyennes de deux pesées.

Cet exemple illustre bien le fait que lorsqu'une certaine latitude est donnée dans la préparation des expériences fournissant des données, des gains peuvent être facilement obtenus.

2 Nécessité de la randomisation

Comme nous allons essayer de vous en convaincre, la randomisation est la seule méthode qui évite les confusions (en fait elle en contrôle la probabilité), qui permet de faire une expérience équitable (ce que l'on pourra prouver), et enfin qui permet

d'apprécier la précision des résultats.

Exemple : Nous voulons comparer l'efficacité de deux médicaments A et B contre la grippe sur 40 malades.

- Expérience 1 : on administre le médicament A aux 20 premiers malades qui se présentent. On note leur état d'amélioration, ensuite on administre le médicament B aux 20 suivants et on note leurs états de santé. A la fin de l'expérience, on calcule les moyennes des états de santé (supposés pouvoir être quantifiés) et on déclare comme meilleur le médicament qui a la meilleure moyenne.
  Cette expérience est désastreuse : (1) durant la durée de l'expérience la maladie, la température extérieure, la fatigue des personnes qui ont réalisé l'expérience ont pu évoluer : l'expérience n'est plus équitable en raison de la confusion possible entre les effets du temps et ceux du médicament ; (2) certains participants de l'expérience, malades ou médecins, qui connaissent parfaitement les médicaments administrés peuvent fausser le résultat inconsciemment : c'est le fameux "effet placebo".

- Expérience 2 : au fur et à mesure qu'un patient arrive en consultation, on alterne strictement A et B, le plan est donc ABABABABABAB... Ce plan est certainement meilleur que le précédent mais il souffre encore de deux gros défauts : (1) la systématicité rend impossible la fameuse méthode dite "double aveugle" (méthode qui est maintenant la norme en médecine et qui requiert que ni les malades ni les médecins ne puissent fausser les résultats de l'expérience), car le médecin, et dans une certaine mesure, le malade, sauront toujours la nature du produit administré ; (2) on ne dispose pas de méthode statistique valide pour choisir entre les situations "A meilleur que B", "B meilleur que A" et "A et B équivalents" ; il y a donc encore risque de confusion.

- Expérience 3 : on pourrait construire une variante de l'expérience 2 dans laquelle on essayerait de répartir au mieux les individus entre deux groupes en fonction de l'âge, du poids, des antécédents. Ce plan, pourtant séduisant, a exactement les mêmes inconvénients que le précédent.

- Expérience 4 : on tire au hasard (tirage équiprobable sans remise) 20 personnes parmi les 40 malades et on leur administre le médicament A. Les autres malades reçoivent le médicament B. Cette expérience (ou plan d'expériences), dite de randomisation totale, possède de nombreuses qualités en regard des précédentes. Afin de préciser et prouver ces propos, formalisons cette expérience.
  En premier lieu, on imagine avoir administré les deux médicaments à chacun des malades et on note R_{ik} la réponse du malade k avec $k = 1, \cdots, n$ (ici $n = 40$) au médicament i avec $i = 1, \cdots, t$ (ici $t = 2$ et $i = 1$ par exemple pour le

médicament A et $i = 2$ pour B). On suppose que le modèle est additif par rapport aux deux effets, qui sont les effets `malade`, notés a_ℓ, et `medicament`, notés m_i, soit :

$$R_{i\ell} = m_i + a_\ell + \varepsilon_{i\ell},$$

avec $\varepsilon_{i\ell}$ une erreur inconnue (nous verrons que cette erreur n'intervient pas sur le traitement du modèle). Pour se débarrasser d'une indétermination, on pose la condition $\sum_{\ell=1}^{n} a_\ell = 0$. **Précisons que la réponse $R_{i\ell}$ est purement conceptuelle et n'est connue qu'en partie** (si elle était totalement connue, l'étude s'arrêterait là).
Maintenant, on note Y_{ij} la réponse du j-ème malade de l'ensemble des malades qui ont reçu le traitement i, sachant que l'on a choisi "au hasard" les n/t (ici 20) malades concernés par le traitement i parmi les n. Du modèle précédent, on déduit que :

$$Y_{ij} = m_i + b_{ij} + \varepsilon_{ij},$$

où l'on peut considérer que les b_{ij} (pour $i = 1, \cdots, t$ et $j = 1, \cdots, n/t$) sont tirés sans remise parmi dans l'ensemble des a_ℓ, $\ell = 1, \cdots, n$. Si l'on procède par ordre lexicographique, b_{11} sera tiré dans l'ensemble $\{a_1, \cdots, a_n\}$, puis b_{12} sera tiré parmi l'ensemble $\{a_1, \cdots, a_n\} \setminus \{b_{11}\}$, etc... **Les réponses Y_{ij} sont donc les réponses effectivement réalisées parmi toutes les $R_{i\ell}$ potentielles**. Cette sélection est aléatoire et ell implique que les b_{ij} sont des effets aléatoires. En utilisant les propriétés d'un tirage uniforme sans remise et la condition $\sum_{\ell=1}^{n} a_\ell = 0$, on montre que les b_{ij} sont des variables aléatoires à valeurs dans $\{a_1, \cdots, a_n\}$ et d'espérance nulle, de matrice de variance-covariance (pour laquelle on a rangé le vecteur $(b_{ij})_{ij}$ en ordre lexicographique) définie par :

$$\begin{aligned} \mathrm{Var}\,(b_{ij}) &= \frac{1}{n}\sum_{\ell=1}^{n} a_\ell^2 \ \text{ pour } 1 \le i \le t,\ 1 \le j \le n/t, \\ \mathrm{cov}\,(b_{ij}, b_{i'j'}) &= -\frac{1}{n(n-1)}\sum_{\ell=1}^{n} a_\ell^2 \ \text{ pour } (i,j) \neq (i',j'). \end{aligned}$$

Pour traiter ce modèle, on effectue maintenant une analyse de la variance sur les Y_{ij} du facteur m_i. Le modèle peut alors s'écrire :

$$Y_{ij} = m_i + b'_{ij} \ \text{ pour } \ 1 \le i \le t,\ 1 \le j \le n/t,$$

où les b'_{ij} sont des termes d'erreur (correspondants à la sommes des b_{ij} et des ε_{ij}) d'espérance nulle et de matrice de variance-covariance presque essentiellement diagonale (les termes extra-diagonaux tous égaux à une constante et petits par

rapport aux termes diagonaux quand n est grand).
Ceci sort un petit peu du cadre de l'analyse de la variance étudiée précédemment (où les erreurs sont non-corrélées, voir chapitre 5), mais du fait que ces termes extra-diagonaux sont constants, on peut tout de même en déduire (voir exercice 4) que leur influence sur les comparaisons entre traitements est nulle. L'analyse de la variance ainsi réalisée pourra, sans confusion, permettre de savoir si les médicaments (effet m_i) sont significativement différents. De plus, on pourra mesurer facilement la précision du résultat obtenu (P-value des tests de Student ou de Fisher par exemple).
Enfin l'expérience 4 permet le travail en "double aveugle" (on suppose que le tirage aléatoire des médicaments a été effectué par un intervenant extérieur). Ceci garantit et prouve le caractère équitable de l'expérience.

3 Plans d'expérience classiques

Dans cette section nous allons présenter les expériences randomisées classiques. Pour le plan en randomisation totale, la démonstration du lien entre la randomisation et le modèle a été donnée ci-dessus. Dans les autres cas, nous admettrons les résultats. Cependant les preuves assez délicates de ces résultats sont données au chapitre 12.

Dans la présentation, nous utiliserons toujours les notations classiques suivantes : l'expérience a pour but principal de comparer t traitements (dans l'exemple médical précédent, $t = 2$) et :

- n est le nombre total d'unités (ou de données) de l'expérience. Dans l'exemple médical, $n = 40$ est le nombre total de malades ;

- r est le nombre de répétitions de chaque traitement. Dans le cas équilibré précédent, $r = n/t$ et en ce qui concerne l'exemple médical, $r = 20$ est le nombre de malades absorbant un des traitements ;

- b est le nombre de blocs. Dans l'exemple précédent, la notion de bloc n'est pas vraiment pertinente. Telle que l'expérience est décrite, un bloc correspondrait à l'ensemble des 40 malades et le nombre de bloc serait $b = 1$.

- k est le nombre d'unités par bloc. Dans l'exemple médical, avec $b = 1$, on a $k = 40$ données par bloc.

3.1 Plan en randomisation totale

Ce type de plan est celui décrit par le plan de l'expérience 4 précédente. Dans le cas général, on a t traitements (le facteur `traitement` admet donc t modalités) , $n = r \cdot t$ unités expérimentales, et on ne considère qu'un bloc ($b = 1$) contenant les $k = n$ unités expérimentales. On tire au hasard (tirage équiprobable et sans remise) r unités pour le premier traitement, puis r pour le second, etc... Les données sont analysées par une analyse de la variance à 1 facteur, le facteur `traitement`. Cependant, en dehors des expériences médicales, ce plan est peu utilisé. On pourrait représenter le plan d'expérience (après randomisation) de l'expérience médicale de l'exemple par le tableau suivant :

Unité(Malade)	Traitement (Médicament)
1	B
2	B
3	A
4	B
5	A
6	A
$\vdots$	$\vdots$
40	A

Construction d'un plan en randomisation totale à l'aide d'un logiciel

Nous allons reprendre l'exemple précédent et construire le plan en randomisation totale correspondant à l'aide des logiciels.

Logiciel SAS :

La solution suivante est relativement acrobatique, la procédure `proc plan` n'est pas vraiment prévue pour ce cas. Nous considérons le cas de 20 unités et 4 traitements.

```
data a;
do unit=1 to 20;
if (unit<=5) then treat=1;
else if (unit<=10) then treat=2;
else if (unit<=15) then treat=3;
else treat=4;
output;
end;
proc plan;
factors unit=20;
output data=a out=b; run;
```

```
proc sort;
by unit; run;
proc print; run;
```

Ici la commande `output` prend le `data a` et randomise son facteur `unit` suivant la randomisation définie par `proc plan`. Cette solution est à rapprocher de la randomisation d'un plan en bloc incomplet. Voici un exemple de plan ainsi crée :

Obs	unit	treat
1	1	4
2	2	1
3	3	2
4	4	4
5	5	3
6	6	3
7	7	2
8	8	2
9	9	3
10	10	2
11	11	1
12	12	1
13	13	2
14	14	4
15	15	3
16	16	3
17	17	1
18	18	4
19	19	1
20	20	4

Logiciel R et Splus :

Voici le même type de création "manuelle" du même plan en randomisation totale, par les logiciels R ou Splus :

```
treat<-sample(rep(1:4,5))
unit<-1:20
plan<-data.frame(unit,treat)
```

Analyse d'un plan en randomisation totale

Voici maintenant l'analyse associée à un tel plan (nous ne donnons ici que les commandes SAS, ce genre d'instructions ayant été traitées également dans les chapitres précédents à l'aide des autres logiciels) :

```
proc glm data=....; /* on peut egalement utiliser proc anova */
class traitement;
model reponse=traitement;
means traitement/tukey; /* a n'examiner que si traitement est significatif*/
run; quit;
```

3.2 Plan en blocs complets (Fisher, 1931)

Ce plan suit les trois principes énoncés par R. Fisher en 1931 : répétition, randomisation et contrôle local. Dans le même cadre que le plan en randomisation totale, on regroupe les $n = r \cdot t$ unités expérimentales en $r = b$ blocs homogènes de taille $k = t$. Dans un exemple médical, les blocs peuvent, par exemple, correspondre au sexe ou à l'âge, dans une expérience agronomique, cela peut être un ensemble de parcelles contiguës, dans toute expérience de laboratoire, cela peut être les unités traitées le même jour, par la même personne.

Le principe du plan est le suivant : dans chaque bloc, on alloue indépendamment une unité à un traitement et ceci de façon aléatoire. Voyons un exemple pour illustrer ce principe.

Exemple 1 : On veut comparer t médicaments à l'aide de $r \cdot t$ rats de laboratoire. On suppose, pour simplifier, qu'ils ont tous le même sexe. Plutôt que de faire un plan en randomisation totale, on décide de les regrouper de manière objective bien qu'un peu arbitraire. Par exemple, on peut les regrouper par poids : les t rats les plus légers constituent le bloc 1, les t suivants le bloc 2 etc... Dans chaque bloc, chaque traitement est administré au hasard à un rat exactement.

Dans l'exemple ci-dessus, les blocs sont constitués de manière arbitraire. Dans certaines situations, la notion de bloc est plus naturelle.

Exemple 2 : On veut comparer plusieurs sels au point de vue de leur parfum. Plus précisément les différents sels sont :

– du sel ordinaire (`sel 1`) ;

– du sel aromatisé aux herbes (`sel 2`) ;

– de la fleur de sel (`sel 3`) ;

– du sel de Guérande (`sel 4`).

Il est conseillé par les vendeurs de sel "haut de gamme" (les deux derniers) de les utiliser crus. On réalise donc l'expérience suivante : le nombre de traitements est $t = 4$, on utilise $r = 3$ tomates qui vont constituer les blocs. De chaque tomate on extrait 4 tranches et on jette les extrémités. On obtient ansi 12 tranches qui constituent les unités de l'expérience.

Il ne reste plus techniquement qu'à : (1) identifier les tranches par un système d'encoches (ce n'est pas très appétissant de coller des étiquettes), (2) faire la randomisation avec la contrainte que les 4 tranches issues d'une même tomate reçoivent les 4 sels différents, (3) faire déguster à l'aveugle. Par exemple une réalisation possible est la suivante :

Tomate	tranche1	tranche 2	tranche 3	tranche 4
1	sel 4	sel 1	sel 2	sel 3
2	sel 3	sel 4	sel 1	sel 2
3	sel 1	sel 3	sel 2	sel 4

De manière générale pour être efficace le plan en blocs complets doit maximiser la variabilité inter-bloc et minimiser la variabilité intra-bloc ; en d'autres termes, il faut rendre l'ensemble des blocs le plus homogène possible. Dans l'exemple sur les rats, on a intérêt à ce que les blocs soient les plus homogènes possible. Par exemple, si on connaît le pedigree des rats, on pourra préférer au critère de poids des critères de parenté pour constituer les blocs. Dans l'exemple des sels, il faut absolument que les 4 tranches issues de la même tomate aient les mêmes caractéristiques géométriques. Ces caractéristiques peuvent varier par contre d'une tomate à l'autre.

Des calculs analogues à ceux fait pour le plan en randomisation totale, mais un peu plus complexes (voir chapitre 12), montrent que l'on peut valider une analyse de la variance à deux facteurs additifs : `traitement` et `bloc`. De plus, cette analyse est équirépétée.

Le plan en blocs complets est quasi-toujours préférable au plan en randomisation totale. C'est le plan le plus employé. C'est celui que l'on essayera d'utiliser a priori.

Construction d'un plan en blocs complets

Logiciel SAS :

Nous supposerons que le lecteur est capable de deviner la logique relativement simple de `proc plan` dans ce cas. Nous prenons l'exemple de 3 blocs et 4 traitements :

```
proc plan;
factors bloc=3 ordered numero=4 ordered;
treatment strt=4 random;
run;
```

Voici un exemple de plan ainsi crée :

Obs	bloc	numero	strt
1	1	1	4
2	1	2	1
3	1	3	3
4	1	4	2
5	2	1	3
6	2	2	2
7	2	3	4
8	2	4	1
9	3	1	1
10	3	2	4
11	3	3	3
12	3	4	2

Logiciel R et Splus :

Voici le même type de création "manuelle" du même plan en blocs complets, par les logiciels R ou Splus :

```
bloc<-sort(rep(1:3,4))
numero<-rep(1:4,3)
strt<-sample(1:4)
for (i in 1:2) {strt<-c(strt,sample(1:4))}
plan<-data.frame(bloc,numero,strt)
```

Analyse d'un plan en blocs complets

Comme précédemment et pour les mêmes raisons, nous ne présentons que le traite-

ment en SAS :

```
proc glm data=...; /* (on peut egalement utiliser proc anova */
class strt bloc;
model reponse=strt bloc;
means strt/tukey /* a n'examiner que si traitement est significatif */;
run;quit;
```

3.3 Plan en blocs incomplets équilibrés ou non

Exemple de plan équilibré : Supposons que nous ayons 9 bières à comparer (le facteur d'intérêt est le facteur `traitement`) et 12 dégustateurs (facteur `bloc`). Comme la bière est un produit amer donc "long en bouche", il est clair que, passé un certain nombre de produits, un dégustateur est incapable de comparer ses sensations. On supposera donc qu'un dégustateur ne peut comparer que 3 bières. Nous avons donc $n = 36$ unités réparties en $b = 12$ blocs de taille $k = 3$ et on propose la répartition suivante :

Blocs	Verre 1	Verre 2	Verre 3
Dégustateur 1	Bière 1	Bière 2	Bière 3
Dégustateur 2	Bière 4	Bière 5	Bière 6
Dégustateur 3	Bière 7	Bière 8	Bière 9
Dégustateur 4	Bière 1	Bière 4	Bière 7
Dégustateur 5	Bière 2	Bière 5	Bière 8
Dégustateur 6	Bière 3	Bière 6	Bière 9
Dégustateur 7	Bière 1	Bière 5	Bière 9
Dégustateur 8	Bière 4	Bière 8	Bière 3
Dégustateur 9	Bière 7	Bière 2	Bière 6
Dégustateur 10	Bière 1	Bière 8	Bière 6
Dégustateur 11	Bière 2	Bière 4	Bière 9
Dégustateur 12	Bière 3	Bière 5	Bière 7

Cette répartition, qui pour l'instant n'a rien d'aléatoire, est équilibrée : chaque traitement se retrouve une fois et une seule exactement avec chaque autre traitement. Plus généralement, un plan en blocs incomplets équilibré est un plan possédant la propriété d'équilibre de la répartition des unités par blocs (comme ci-dessus), c'est-à-dire que tous les traitement sont appliqués au total le même nombre de fois et que deux traitements se retrouvent ensemble dans un bloc le même nombre de fois. On peut donc décrire (partiellement) un plan en blocs incomplets équilibré par :

– le nombre de traitements t (9 dans l'exemple) ;

- le nombre de répétitions r (4 dans l'exemple) ;
- le nombre de blocs b (12 dans l'exemple) ;
- la longueur d'un bloc k (3 dans l'exemple) ;
- l'indice de concurrence λ : le nombre de fois où deux traitements se retrouvent ensemble dans un bloc (1 dans l'exemple).

En comptant de deux manières différentes le nombre de parcelles et le nombre de voisins, on obtient :

$$r \cdot t = b \cdot k \quad \text{et} \quad r(k-1) = \lambda(t-1).$$

Cependant, il n'existe pas des plans pour toutes tailles (nombre de traitements, de blocs,...) et qui vérifient ces équations. On pourra consulter à ce sujet des tables de plans (voir par exemple Raghavarao [49]). Quand il existe un plan équilibré, on montre qu'il est optimal (pour plus de détails, voir Druilhet [25]). Malheureusement cela n'est pas toujours possible. En particulier, l'équilibre demande souvent un nombre de répétitions élevé. Dans ce cas, il existe des méthodes pour construire des plans conservant certaines propriétés, par exemple l'équilibre partiel (voir Coursol [19]). De toutes manières, les propriétés suivantes restent vraies que le plan soit équilibré ou non :

i. **Randomisation :** la randomisation se fait en deux étapes qui sont

   1. le "mélange des blocs" : si on reprend l'exemple précédent, on affecte un numéro de dégustateur à un dégustateur réel (M. Dupond) au hasard ;
   2. le "mélange des traitements par bloc" : toujours en reprenant l'exemple, les trois bières devant être présentées à un dégustateur le sont dans un ordre aléatoire.

ii. **Analyse :** on montre que la randomisation valide un modèle possédant des effets `traitement` fixes et des effets `bloc` aléatoires : c'est un modèle mixte. Si on ne dispose pas des moyens de traiter un tel modèle, on peut toutefois utiliser un modèle avec blocs et traitements fixes qui correspond à une légère perte d'information.

Ainsi, dans l'exemple des bières, pourra-t-on se retrouver avec le plan à blocs incomplet équilibré suivant :

Blocs	Essai 1	Essai 2	Essai 3
Dégustateur 8	Bière 3	Bière 1	Bière 2
Dégustateur 1	Bière 4	Bière 5	Bière 6
Dégustateur 7	Bière 8	Bière 7	Bière 9
Dégustateur 12	Bière 1	Bière 4	Bière 7
Dégustateur 10	Bière 8	Bière 2	Bière 5
Dégustateur 9	Bière 6	Bière 9	Bière 3
Dégustateur 5	Bière 1	Bière 9	Bière 5
Dégustateur 2	Bière 8	Bière 4	Bière 3
Dégustateur 11	Bière 2	Bière 7	Bière 6
Dégustateur 4	Bière 1	Bière 8	Bière 6
Dégustateur 6	Bière 2	Bière 9	Bière 4
Dégustateur 3	Bière 7	Bière 5	Bière 3

Pour arriver à obtenir un plan avant randomisation comme celui de cet exemple, il existe différentes méthodes classiques. Voyons donc les principaux types de plan à blocs incomplets équilibrés ou non.

Les plans lattices :

En premier lieu, voici une méthode pour construire des plans qui sont, sous certaines conditions, équilibrés, et qui possèdent de toutes façons de bonnes propriétés : ce sont les plans lattices. Cependant, avant d'aller plus avant, un rappel d'algèbre est nécessaire :

Tout d'abord, on rappelle qu'un corps est un ensemble K, muni de deux lois $+$ et $\cdot$ dans lequel on peut toujours résoudre une équation du type $ax + b = 0$. Les exemples classiques de corps sont : $\mathbb{Q}$, corps des nombres rationnels, $\mathbb{R}$, corps des nombres réels et $\mathbb{C}$, corps des nombres complexes.
Ici, ce sont plutôt les corps finis qui nous intéressent. On connaît un premier exemple de corps fini $\mathbb{F}_p$ de cardinal p : dans le cas où p est un nombre premier, alors $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, corps des entiers modulo p. Ce corps est unique à un isomorphisme (application bijective respectant addition et multiplication) près. Plus généralement, on a les propriétés suivantes :

Théorème 11.1 *Soit K un corps fini.*

i. Le nombre d'éléments $Card(K) = q$, est une puissance d'un nombre premier : $q = p^r$.

ii. Il existe toujours un corps de cardinal $q = p^r$ et concrètement il n'en existe qu'un seul. Plus mathématiquement : deux corps finis de même cardinal sont

isomorphes.

iii. Pour tout nombre premier p et tout entier r, le corps de cardinal $q = p^r$ peut être décrit comme l'extension galoisienne de $\mathbb{F}_p$ modulo un polynôme irréductible de degré r.

Par exemple, pour $q = 4 = 2^2$, le polynôme $X^2 + X + \bar{1}$ est irréductible dans $\mathbb{F}_2$, ce qui veut dire qu'il n'a pas de racines. En effet $\bar{0}^2 + \bar{0} + \bar{1} = \bar{1} \neq \bar{0}$ et $\bar{1}^2 + \bar{1} + \bar{1} = \bar{1} \neq \bar{0}$. De la même façon que pour la construction de $\mathbb{C}$, on note x une racine "imaginaire" de ce polynôme. Il est facile de voir que l'on obtient ainsi 4 élément de $\mathbb{F}_4$: $(\bar{0}, \bar{1}, x, x+\bar{1})$.

Notons que pour tout corps fini $\mathbb{F}_p$, $\bar{0}$ et $\bar{1}$ appartiennent à ce corps.
Nous allons utiliser cette notion pour définir un plan lattice :

Définition 11.1 *Un (r, p^2) lattice est un plan pour p^2 traitements (avec p un nombre premier ou une puissance d'un nombre premier) avec r répétitions de p blocs contenant chacun p unités (donc au total il y a $r \cdot p$ blocs). Aux p^2 traitements sont associés deux facteurs A et B à p niveaux, ces niveaux étant numérotés par des éléments du corps de Galois $\mathbb{F}_p$ à p éléments. Un traitement est donc indicé par une paire (i, j) à valeur dans $\mathbb{F}_p^2$, i est la valeur prise par le facteur A, j est la valeur prise par le traitement B. Les répétitions sont indicées par $\mathbb{F}_p \cup \{\infty\}$, les blocs dans chaque répétition sont indicés par $k \in \mathbb{F}_p$:*

i. lors de la répétition ∞, on affecte au bloc $k \in \mathbb{F}_p$ les traitements (i, j) tels que $j = k$.
On dit que l'on a confondu le facteur bloc avec le facteur B.

ii. lors de la répétition $\bar{0}$, on affecte au le bloc $k \in \mathbb{F}_p$ les traitements (i, j) tels que $i = k$.
On dit que l'on a confondu le facteur bloc avec le facteur A.

iii. lors de la répétition $\ell \in \mathbb{F}_p \setminus \{0\}$, on affecte au bloc $k \in \mathbb{F}_p$ les traitements (i, j) tels que $i + \ell \cdot j = k$.
On dit que l'on a confondu le facteur bloc avec le facteur $A + \ell \cdot B$.

On montre qu'un $((p + 1), p^2)$ lattice est un plan équilibré. Illustrons ceci par des exemples :

Exemples de plan lattice : • un $(3, 2^2)$ lattice est un plan pour 4 traitements. Chaque traitement est indicé arbitrairement par deux facteurs A et B prenant chacun deux modalités numérotés dans $\mathbb{F}_2$, corps de Galois à deux éléments, c'est-à-dire

$\mathbb{Z}/2\mathbb{Z}$.
Ces 4 traitements sont expérimentés sur 12 unités divisées en 3 répétitions de deux blocs de taille deux.

Blocs	Trait. dans unité 1	Trait. dans unité 2
Bloc 1 = répétition ∞, bloc $\bar{0}$	$(\bar{0},\bar{0}) = \mathbf{1}$	$(\bar{1},\bar{0}) = \mathbf{3}$
Bloc 2 = répétition ∞, bloc $\bar{1}$	$(\bar{0},\bar{1}) = \mathbf{2}$	$(\bar{1},\bar{1}) = \mathbf{4}$
Bloc 3 = répétition $\bar{0}$, bloc $\bar{0}$	$(\bar{0},\bar{0}) = \mathbf{1}$	$(\bar{0},\bar{1}) = \mathbf{2}$
Bloc 4 = répétition $\bar{0}$, bloc $\bar{1}$	$(\bar{1},\bar{0}) = \mathbf{3}$	$(\bar{1},\bar{1}) = \mathbf{4}$
Bloc 5 = répétition $\bar{1}$, bloc $\bar{0}$	$(\bar{0},\bar{0}) = \mathbf{1}$	$(\bar{1},\bar{1}) = \mathbf{4}$
Bloc 6 = répétition $\bar{1}$, bloc $\bar{1}$	$(\bar{1},\bar{0}) = \mathbf{3}$	$(\bar{0},\bar{1}) = \mathbf{2}$

$\xrightarrow[\text{(randomisation)}]{\Longrightarrow}$

Blocs	Trait. Un. 1	Trait. Un.2
Bloc 1 (ex 3)	**1**	**3**
Bloc 2 (ex 1)	**2**	**1**
Bloc 3 (ex 6)	**3**	**2**
Bloc 4 (ex 2)	**4**	**3**
Bloc 5 (ex 5)	**1**	**4**
Bloc 6 (ex 4)	**4**	**2**

(par exemple).

• de la même manière, un $(5, 4^2)$ lattice est un plan comportant 16 traitements différents, et 5 répétitions de 4 blocs, chaque bloc contenant 4 unités (ce qui fait un total de 80 unités statistiques). On utilise le corps $\mathbb{F}_4 = \{\bar{0}, \bar{1}, x, \bar{1} + x\}$ pour construire un tel plan (on a dans ce corps, par exemple, pour l'addition, $\bar{1}+(\bar{1}+x) = x$, $x + (\bar{1} + x) = \bar{1}$ ou $(\bar{1} + x) + (\bar{1} + x) = \bar{0}$ et pour la multiplication, $x \cdot (\bar{1} + x) = \bar{1}$ ou $(\bar{1} + x) \cdot (\bar{1} + x) = x$) car $\bar{1} + x + x^2 = \bar{0}$. Voici quelques lignes du tableau avant randomisation :

Blocs	Trait. Un. 1	Trait. Un.2	Trait. Un.3	Trait. Un. 4
Bloc 1 = répétition ∞, bloc $\bar{0}$	$(\bar{0},\bar{0})$	$(\bar{1},\bar{0})$	$(x,\bar{0})$	$(\bar{1}+x,\bar{0})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Bloc 11 = répétition $\bar{1}$, bloc x	$(\bar{0},x)$	$(\bar{1},1+x)$	$(x,\bar{0})$	$(\bar{1}+x,\bar{1})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Bloc 18 = répétition $\bar{1}+x$, bloc $\bar{1}$	$(\bar{0},x)$	$(\bar{1},\bar{0})$	$(x,\bar{1})$	$(\bar{1}+x,\bar{1}+x)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Les plans circulants :

Voici maintenant un autre type de plan en blocs incomplets équilibré, lorsque l'on dispose de $t(t-1)$ unités pour t traitements :

Définition 11.2 (Plans circulants) *Un plan circulant est un plan pour t traitements en t blocs de tailles $t-1$. Pour construire les blocs, on élimine tour à tour*

chacun des traitements. Un plan circulant est toujours équilibré, son indice de concurrence λ vaut $t-2$.

Exemple de plan circulant : Si on considère un plan circulant à 5 traitements (notés $t_1, \cdots, t_5$), on a 5 blocs de taille 4 dans lesquels on élimine tour à tour un des traitement. Par exemple :

Blocs	Unité 1	Unité 2	Unité 3	Unité 4
Bloc 1	t_2	t_3	t_4	t_5
Bloc 2	t_1	t_3	t_4	t_5
Bloc 3	t_1	t_2	t_4	t_5
Bloc 4	t_1	t_2	t_3	t_5
Bloc 5	t_1	t_2	t_3	t_4

$\Longrightarrow$ (randomisation)

Blocs	Unité 1	Unité 2	Unité 3	Unité 4
Bloc 1 (ex 3)	t_4	t_5	t_1	t_2
Bloc 2 (ex 2)	t_5	t_3	t_1	t_4
Bloc 3 (ex 5)	t_4	t_2	t_3	t_1
Bloc 4 (ex 4)	t_2	t_1	t_5	t_3
Bloc 5 (ex 1)	t_4	t_5	t_3	t_2

Construction d'un plan en blocs incomplets à l'aide d'un logiciel

De manière générale, il n'y a pas de solution informatique dans tous les cas. Une solution consiste à construire un plan initial non randomisé à partir de méthodes algébriques ou à partir de tables de plans, que l'on randomise par la suite.

Logiciel SAS :

Encore une fois, la commande `proc plan` va être au fondement de la construction (voir l'exemple suivant), mais on utilise également la commande `output data=entree out=sortie`, qui permet de lire une table `entree` contenant un plan, de le randomiser (en utilisant `proc plan`) et de le sortir sur une table `sortie`. Considérons par exemple un plan en 2 blocs de 3 sous-blocs de 3 parcelles, par exemple un plan lattice pour 9 traitements. La randomisation se fera par :

```
proc plan;
factors bloc=2 sbloc=3 par=3;
output out=sortie;
run;
proc print;
run;
```

```
proc sort data=sortie;
by bloc sbloc par;
proc print;
run; quit;
```

d'où le résultat suivant :

```
 bloc    sbloc      -par-
 2          3       3 1 2
            2       3 2 1
            1       3 2 1
 1          3       3 1 2
            1       2 3 1
            2       1 2 3

-------------------------------------
Obs    bloc    sbloc    par
 1      1       2       2
 2      1       2       1
 3      1       2       3
 4      1       1       2
 5      1       1       1
 :      :       :       :
```

Logiciel R et Splus :

Voici le même type de création "manuelle" du même plan en blocs incomplets, par les logiciels R ou Splus :

```
blo<-sample(1:2)
bloc<-c(rep(blo[1],9),rep(blo[2],9))
sblo<-c(sample(1:3),sample(1:3))
sbloc<-c(rep(sblo[1:3],3),rep(sblo[4:6],3))
par<-sample(1:3)
for (i in 1:5) {par<-c(par,sample(1:3))}
plan<-data.frame(bloc,sbloc,par)
```

Deux cas particuliers méritent notre attention : les plans circulants et les plans lattices.

Exemple de construction d'un plan circulant

Dans le premier cas (plans circulants), voici un exemple dans lequel on désire appliquer 6 traitements à 5 parcelles :

Logiciel SAS :

```
proc plan;
factors bloc=6 parcelle=5;
treatment trt=5 of 6 cyclic (1 2 3 4 5);
run;quit;
```

Voici un résultat possible :

```
bloc      ---parcelle--      ---trt---
   4      4  5  1  2  3      1 2 3 4 5
   5      1  2  5  3  4      2 3 4 5 6
   1      2  5  1  3  4      3 4 5 6 1
   6      4  1  2  3  5      4 5 6 1 2
   2      1  5  4  3  2      5 6 1 2 3
   3      5  4  2  1  3      6 1 2 3 4
```

Logiciel R et Splus :

Voici à nouveau un bricolage permettant d'obtenir le même plan d'expérience en R ou Splus :

```
A<-rep(c(1:6),2)
trt<-matrix(1:30,6)
for (i in c(1:6)) {trt[i,1:5]<-A[i:(i+4)]}
par<-sample(1:5)
for (i in 1:5) {par<-c(par,sample(1:5))}
parcelle<-matrix(par,6)
blo<-sample(c(1:6))
for (i in c(1:6)) {bloc[(5*i-4):(5*i)]<-blo[i]}
trt<-as.vector(trt)
parcelle<-as.vector(parcelle)
plan<-data.frame(bloc,parcelle,trt)
```

Exemple de construction d'un plan lattice

Là encore, nous donnons la solution sans commentaire en reprenant l'exemple de dégustation de bières, mais seulement pour le logiciel SAS.

Logiciel SAS :

Pour les plans lattices, il faut utiliser la commande `proc factex` qui est incluse dans le module contrôle de qualité.

```
proc factex;
factors x1-x4/nlev=3;
size design=9;
model r=3;
output out=a x1 nvals=(0 1 2)
             x2 nvals=(0 1 2)
             x3 nvals=(0 1 2)
             x4 nvals=(0 1 2);
run;
proc print data=a;run;

data b;
keep rep block plot t;
array x{4} x1-x4;
do rep=1 to 4;
    do block=1 to 3;
        plot=0;
            do n=1 to 9;
            set a point=n;
            if (x{rep}=block-1) then do;
                t=n;
                plot=plot+1;
                output;
            end;
        end;
    end;
end;
stop;
run;

data _null_;
array s{3} s1-s3;
```

```
file print;
n=1;
    do r=1 to 4;
    put "Replication" r 1.0 ":";
        do b=1 to 3;
            do p=1 to 3;
                set b point=n;
                s{plot}=t;
                n=n+1;
            end;
        put "Degustateur " b 1.0 ":" (s1-s3) (3.0);
        end;
    put;
    end;
stop;
run;
```

Voici un résultat possible :

```
Obs    x1    x2    x3    x4
 1      0     0     0     0
 2      0     1     2     2
 3      0     2     1     1
 4      1     0     2     1
 5      1     1     1     0
 6      1     2     0     2
 7      2     0     1     2
 8      2     1     0     1
 9      2     2     2     0

Replication 1:
 Degustateur 1:  1  2  3
 Degustateur 2:  4  5  6
 Degustateur 3:  7  8  9

Replication 2:
 Degustateur 1:  1  4  7
 Degustateur 2:  2  5  8
 Degustateur 3:  3  6  9

Replication 3:
```

```
 Degustateur 1:  1  6  8
 Degustateur 2:  3  5  7
 Degustateur 3:  2  4  9

Replication 4:
 Degustateur 1:  1  5  9
 Degustateur 2:  3  4  8
 Degustateur 3:  2  6  7
```

Analyse d'un plan en blocs incomplets

```
proc mixed data=...;
class traitement bloc ;
model reponse=traitement;
random bloc;
lmeans traitement/tdiff /* a n'examiner que si traitement significatif
                           il faut utiliser la methode de Bonferroni */;
run;quit;
```

Dans les plans en blocs complets ou incomplets, les unités sont regroupées selon une classification : les blocs. Dans certains autres cas, les unités peuvent être rangées dans une structure en lignes et colonnes ; ceci fera l'objet de la prochaine section.

3.4 Plans en lignes et colonnes

Cas équilibré : les carrés latins

Définition 11.3 *On appelle carré latin, un plan comprenant n^2 unités, pour trois facteurs ayant le même nombre n de niveaux, et tel que le nombre de répétitions d'une paire de niveaux pour deux facteurs est toujours de* 1.

Exemple : on veut comparer 4 peintures sur 4 maisons carrées ayant 4 façades de mêmes expositions N, S, E, O. Le facteur d'intérêt est le facteur `peinture`, les facteurs `maison` et `orientation` étant des facteurs à contrôler type bloc. Si on veut équilibrer les relations du premier facteur (`peinture`) avec chacun des deux autres (`maison` et `orientation`), on peut être amené (avant randomisation) à utiliser la répartition suivante :

	Orientation			
Maison	Nord	Sud	Est	Ouest
1	peinture 1	peinture 2	peinture 3	peinture 4
2	peinture 2	peinture 3	peinture 4	peinture 1
3	peinture 3	peinture 4	peinture 1	peinture 2
4	peinture 4	peinture 1	peinture 2	peinture 3

On satisfera à l'exigence de randomisation en faisant dans l'ordre que l'on voudra un "mélange des lignes" et "un mélange des colonnes". Cette randomisation valide, sous les mêmes hypothèses d'additivité que pour le plan en randomisation totale, une analyse de la variance à trois facteurs additifs. On pourra par exemple obtenir, après mélange des lignes :

	Orientation			
Maison)	Nord	Sud	Est	Ouest
1	peinture 2	peinture 3	peinture 4	peinture 1
2	peinture 3	peinture 4	peinture 1	peinture 2
3	peinture 4	peinture 1	peinture 2	peinture 3
4	peinture 1	peinture 2	peinture 3	peinture 4

Puis, après mélange des colonnes,

	Orientation			
Maison)	Nord	Sud	Est	Ouest
1	peinture 4	peinture 3	peinture 1	peinture 2
2	peinture 2	peinture 1	peinture 3	peinture 4
3	peinture 1	peinture 4	peinture 2	peinture 3
4	peinture 3	peinture 2	peinture 4	peinture 1

Remarquez que les propriétés combinatoires sont conservées.

La méthode de construction "en diagonale" fonctionne pour toute valeur de n. Mais il existe de nombreuses méthodes de construction. La plus classique est basée sur un groupe fini. On construit la valeur du troisième facteur comme la table d'addition de deux permutations du groupe (noté additivement).

Construction et randomisation d'un carré latin

Logiciel SAS :

Les commandes SAS pour obtenir un tel plan sont :

```
proc plan;
factors ligne=6 ordered colonne=6 ordered;
treatment trt=6 cyclic (6 1 5 2 4 3);
output out=a
ligne random
colonne random;
output out=b;
run;quit;
```

Voici un résultat possible :

```
Factor       Select      Levels     Order
ligne             6           6    Ordered
colonne           6           6    Ordered

Treatment Factors

Initial Block
Factor      Select      Levels     Order     / Increment
trt              6           6    Cyclic     (6 1 5 2 4 3) / 1

 ligne      --colonne--      ----trt----
 1          1 2 3 4 5 6      6 1 5 2 4 3
 2          1 2 3 4 5 6      1 2 6 3 5 4
 3          1 2 3 4 5 6      2 3 1 4 6 5
 4          1 2 3 4 5 6      3 4 2 5 1 6
 5          1 2 3 4 5 6      4 5 3 6 2 1
 6          1 2 3 4 5 6      5 6 4 1 3 2

           Obs    ligne    colonne   trt
           1      4         2        6
           2      4         3        1
           3      4         6        5
           4      4         5        2
           5      4         1        4
           6      4         4        3
           7      1         2        1
           8      1         3        2
           9      1         6        6
          10      1         5        3
          11      1         1        5
```

```
12      1       4       4
13      2       2       2
:       :       :       :
```

Notez que le bloc initial qui est particulier donne au plan des propriétés d'équilibre des voisinages en ligne. Si on ne le précise pas, il sera par défaut 1 2 3 4 5 6.

Logiciel R et Splus :

Voici la génération d'un carré latin (particulier) qui reprend celle de la matrice circulante précédente. Ce traitement est donc moins général que celui donné en SAS :

```
A<-rep(c(1:6),2)
trt<-matrix(1:36,6)
for (i in c(1:6)) {trt[i,1:6]<-A[i:(i+5)]}
lig<-sample(1:6)
col<-sample(1:6)
trt[lig,col]<-trt
traitement<-as.vector(t(trt))
obs<-c(1:36)
plan<-data.frame(obs,traitement)
```

Analyse d'un carré latin

Pour analyser un carré latin avec le logiciel SAS, on utilise à nouveau la procédure `proc glm` :

```
proc glm data=...; /* on peut egalement utiliser proc anova */
class traitement ligne colonne;
model reponse=traitement ligne colonne;
means traitement/tukey: /* a n'examiner que si traitement significatif;
run;quit;
```

Les plans en lignes et colonnes équilibrés ou non :

On peut s'intéresser à un plan dont les unités sont placées sur un réseau de l lignes et de c colonnes mais dont l'allocation des traitements n'a plus la belle propriété du carré latin. On utilise encore la même randomisation, mais elle valide maintenant une analyse avec un effet `ligne` et un effet `colonne` aléatoires.

Logiciel SAS :

```
proc mixed data=...;
class traitement ligne colonne;
model reponse=traitement;
random ligne colonne;
lsmeans traitement/tdiff /* a n'examiner que si traitement significatif
                            il faut utiliser la methode de Bonferroni */;
run;quit;
```

3.5 Les plans split-plot (parcelle subdivisée) :

Exemple : Simplifions la présentation d'une expérience célèbre de Cochran et Cox [18]. On désire comparer 3 recettes de gâteau au chocolat et 6 températures de cuisson $180, 190, \cdots, 230$ degrés. On réalise l'expérience suivante : chaque jour (pendant 4 jours), on réalise 3 pâtes correspondant aux 3 recettes. Chaque pâte est ensuite sub-divisée en 6 sous-parties cuites à chacune des 6 températures différentes. Après cuisson, on mesure le moelleux du gâteau en mesurant l'angle de rupture α d'une tranche. On a donc pour expliquer cette variable quantitative trois facteurs : la température (notée `temperature`, que l'on décide de considérer comme qualitatif), la recette (notée `recette`) et le jour (facteur `bloc`). Ce qui est particulier à cet exemple est qu'il y a deux types d'erreurs : une erreur attachée à la mesure de l'angle et l'autre à la confection d'une pâte.

On aura ainsi par exemple la planification suivante :

i. on dresse d'abord un plan en blocs complets randomisé pour le facteur `recette`, par exemple :

Blocs (jours)			
Bloc 1	Rec.2	Rec.1	Rec.3
Bloc 2	Rec.3	Rec.2	Rec.1
Bloc 3	Rec.2	Rec.3	Rec.1
Bloc 4	Rec.1	Rec.3	Rec.2

ii. ensuite, chaque "grande parcelle" [1] (ici les différentes pâtes de gâteau), est divisée en autant de "sous-parcelles" qu'il y a de niveaux du second facteur traitement : la température. Ce dernier facteur est pris dans un ordre randomisé.

1. A l'origine ces plans étaient des expérimentations agronomiques dont les unités étaient des parcelles d'ou le vocabulaire qui est passé dans l'usage

Une réalisation possible peut être :

Blocs (jours)			
Bloc 1	Rec.2	Rec.1	Rec.3
	$T_4, T_2, T_1, T_6, T_5, T_3$	$T_4, T_5, T_6, T_2, T_3, T_1$	$T_1, T_3, T_2, T_4, T_5, T_3$
Bloc 2	Rec.3	Rec.2	Rec.1
	$T_3, T_6, T_1, T_5, T_4, T_2$	$T_4, T_1, T_2, T_6, T_5, T_3$	$T_6, T_2, T_1, T_4, T_3, T_5$
Bloc 3	Rec.2	Rec.3	Rec.1
	$T_2, T_3, T_5, T_4, T_1, T_6$	$T_2, T_6, T_3, T_4, T_5, T_1$	$T_5, T_1, T_3, T_2, T_6, T_4$
Bloc 4	Rec.1	Rec.3	Rec.2
	$T_1, T_4, T_2, T_3, T_5, T_6$	$T_4, T_5, T_6, T_2, T_1, T_3$	$T_3, T_4, T_1, T_2, T_6, T_5$

Plus généralement, on a la définition suivante :

Définition 11.4 *On appelle plan split-plot un plan pour deux facteurs traitements, le premier traitement, A, possédant t niveaux et le second ,B, s niveaux. Pour le premier facteur, on construit un plan en blocs complets à $r \cdot t$ "grandes unités" avec sa randomisation. Ensuite, chaque "grande unité" est divisée en s sous-unités auxquelles sont affectées dans un ordre aléatoire les s valeurs du traitement B.*

Si i, j, k sont, respectivement, les niveaux de A, B et *bloc*, le modèle auquel conduit la randomisation est

$$Y_{ijk} = \mu + a_i + b_j + (a*b)_{ij} + bloc_k + E_{ik} + \varepsilon_{ijk} \text{ pour } i = 1, \cdots, t,\ j = 1, \cdots, s \text{ et } k = 1, \cdots, r,$$

avec E l'erreur associée aux "grandes unités" (dans l'exemple, Y désigne la mesure de l'angle α d'une tranche, A est le facteur `temperature`, B est le facteur `recette` et le facteur `bloc` correspond à un jour).

Ce modèle peut s'analyser bien évidemment à l'aide d'un programme de modèle mixte, mais il peut également s'analyser à partir des projections inter- et intra- "grande unité". La projection inter- revient à travailler sur $Y_{i.k}$, ce qui, avec des contraintes classiques, donne

$$Y_{i.k} = \mu + a_i + bloc_k + \varepsilon'_{ik} \quad \text{avec} \quad \varepsilon'_{ik} = E_{ik} + \varepsilon_{i.k}.$$

Les deux aléas peuvent être confondus, on retrouve ainsi le modèle du plan en blocs complets. La projection intra- est théoriquement basée sur les $Y_{ijk} - Y_{i.k}$. On montre qu'elle est équivalente au modèle complet dans lequel l'effet E_{jk} est supposé fixe, ceci à condition de se limiter à l'estimation et aux tests sur le facteur B et sur l'interaction $A * B$.

En conclusion on dit que le facteur A est totalement estimable inter-grandes unités et l'interaction $A * B$ et le facteur B sont totalement estimables intra-grandes unités.

De manière générale, le plan split-plot donne une meilleure précision sur le second facteur B et sur l'interaction $A * B$ que sur le premier facteur. Le lecteur pourra le vérifier facilement sur un exemple.

Construction, randomisation et analyse d'un plan split-plot

Logiciel SAS :

Voici les commandes permettant de construire un tel plan :

```
proc plan;
factors bloc=3 ordered parcelle=4 soup=5;
run;quit;
```

Logiciel R et Splus :

Voyons maintenant la génération du même plan avec les logiciels R et Splus :

```
blo<-sample(1:3)
bloc<-c(rep(blo[1],20),rep(blo[2],20),rep(blo[3],20))
par<-c(sample(1:4),sample(1:4),sample(1:4))
parc<-c(rep(par[1:4],5),rep(par[5:8],5),rep(par[9:12],5))
soup<-sample(1:5)
for (i in 1:11) {soup<-c(soup,sample(1:5))}
plan<-data.frame(bloc,parc,soup)
```

Voici maintenant en SAS les commandes à utiliser pour analyser un plan split-plot :

```
proc mixed data=...;
class trait1 trait2 bloc;
model reponse=trait1 trait2 trait1*trait2 bloc;
random trait1*bloc;
run;quit;
```

4 Exercices

Exercice 11.1

(**) On reprend les termes de l'expérience 4 décrite dans la section 2. Prouver alors l'affirmation :

$$\mathrm{Var}\left(Y_{i.} - Y_{i'.}\right) = \frac{2}{n.r}\sum_{k=1}^{n} a_k^2 \quad \text{où} \quad i \neq i' \quad \text{et} \quad r = n/I.$$

En déduire les conclusions obtenues sur l'analyse de la variance de l'expérience 4.

Exercice 11.2

(*) En utilisant la description de F_3 comme l'ensemble $\{\bar{0}, \bar{1}, \bar{2}\}$ muni de l'addition et de la multiplication modulo 3, construire le lattice $(4, 3^2)$.

Exercice 11.3

(**) En utilisant la description de F_4 comme l'ensemble $\{\bar{0}, \bar{1}, x, \bar{1} + x\}$ muni de l'addition modulo 2, de la multiplication modulo 2 et du polynôme irréductible $\bar{1} + x + x^2$, construire le lattice $(5, 4^2)$. Par exemple, sur F_4, $(1+x)(1+x) = \bar{1} + 2x + x^2 = x$

Exercice 11.4

(*) Vérifiez informatiquement que pour un carré latin on peut déclarer indifféremment les effets lignes et colonnes comme fixes ou aléatoires (cela ne change pas les résultats).

Exercice 11.5

(***) [Carrés latins complets en ligne et en colonnes] Dans une expérience qui a une structure spatiale et dont les unités sont disposées dans un réseau en lignes et colonnes, les voisinages peuvent avoir un intérêt. C'est la cas par exemple quand on soupçonne des contaminations entre parcelles voisines dans des expériences agronomiques. On peut par exemple désirer équilibrer les voisinages : chaque paire de traitement doit apparaître exactement le même nombre de fois comme voisins. Voici la solution la plus simple à ce problème. Soit $\mathbb{Z}/n\mathbb{Z}$ le groupe cyclique des entiers modulo n. Il s'agit de $\{0, 1, \cdots, n-1\}$ muni de l'addition modulo n (par exemple pour $n = 5$, on a $4 + 3 = 2$). Supposons n pair et considérons la suite de Williams de taille n : $w_1, \cdots, w_n := 0, 1, n-1, 2, n-2, \cdots$.

i. Montrer que le dernier élément vaut $n/2$.

ii. Montrer que $w_1, \cdots, w_n$ est une permutation de $\mathbb{Z}/n\mathbb{Z}$.

iii. Montrer que l'ensemble $\{w_{i+1} - w_i,\ i = 1, \cdots, n-1\}$ donne une occurrence exactement de tout élément non nul de $\mathbb{Z}/n\mathbb{Z}$.

iv. Vérifier que pour $n = 6$, la table d'addition de $w_1, \cdots, w_6$ par lui même (c'est-à-dire le tableau à double entrée $(w_i + w_j)_{1 \leq i,j \leq n}$ donne un carré latin équilibré pour les voisinages.

v. Montrer cette propriété pour tout n pair.

Pour en savoir plus on pourra consulter Azaïs *et al.* [4].

Chapitre 12

Plans randomisés par un groupe de permutations : la théorie

Ce chapitre énonce les résultats théoriques légitimant la construction des plans d'expériences randomisés présentés dans le chapitre précédent.

1 Le modèle de la randomisation

Ce chapitre donne les bases théoriques permettant de mieux comprendre comment la randomisation permet de valider le modèle d'analyse des plans vus au chapitre 11. Comme les situations rencontrées sont assez diverses, nous allons nous placer dans un cadre abstrait très général qui les englobera toutes.

1.1 Présentation générale

Considérons une expérience qui comprend n unités expérimentales. On note Ω l'ensemble de ces unités que l'on peut identifier à $\{1, \cdots, n\} = \Omega$. On suppose que cette expérience a pour but de comparer un ensemble $T = \{1, \cdots, t\}$ de traitements, avec la contrainte que chaque unité ne peut recevoir qu'un seul traitement. On considère un "plan initial" qui correspond à un premier placement des traitements dans les unités, c'est-à-dire à une application de Ω dans T et qui peut se représenter par une matrice binaire (avec uniquement des 0 et des 1)de taille (n, t), que nous noterons X.

Exemple : Considérons le cas où $n = 5$ et $t = 3$ et le plan initial $1, 2, 3, 1, 2$, c'est-à-dire que l'unité 1 reçoit le traitement 1, puis l'unité 2 le traitement 2,..., et enfin, l'unité 5 le traitement 2. Alors la matrice X est la suivante :

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Avant de poursuivre, il nous faut revenir sur quelques notions algébriques qui vont s'avérer utiles par la suite.

1.2 Quelques notions d'algèbre

On rappelle d'abord qu'un groupe est un ensemble dans lequel les éléments vérifient des propriétés de stabilité, d'associativité d'existence d'un élément neutre et d'un symétrique pour une certaine opération que l'on notera + (par exemple, $\mathbb{Z}/3\mathbb{Z}$ avec l'opération +).

On note S_n le groupe des permutations de $\{1, \cdots, n\}$ (c'est-à-dire l'ensemble des bijections de $\{1, \cdots, n\}$ dans $\{1, \cdots, n\}$) auquel on associe l'opération de composition entre applications.

Définition 12.1 *Un groupe $(G, .)$ inclus dans S_n est simplement transitif sur Ω si*

$$\forall x, y \in \Omega, \quad \exists \xi \in G \text{ tel que } \xi(x) = y.$$

S_n est évidemment un groupe simplement transitif sur Ω. Soit maintenant H un sous-groupe simplement transitif de S_n. Par exemple, pour $n = 3$, H peut être constitué des 3 permutations :

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}.$$

Tout élément ξ de H est identifié à une matrice que l'on notera encore ξ et qui effectue la permutation correspondante des coordonnées. Par exemple, à la permutation ξ suivante de S_5,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 5 & 3 \end{pmatrix},$$

on associe la matrice

$$\xi = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

Définition 12.2 *Pour H un sous-groupe de S_n simplement transitif sur $\Omega = \{1, \cdots, n\}$, l'orbite double de la paire d'élément $(i, j) \in \Omega^2$ est l'ensemble*

$$O_{ij} := \big\{\big(\xi(i), \xi(j)\big) : \xi \in H\big\}.$$

A l'orbite O_{ij}, on associe la matrice $M = (m_{i',j'})_{1 \leq i',j' \leq n}$ (qui dépend elle aussi de (i, j) mais nous l'omettrons pour simplifier les notations) telle que :

$$m_{i'j'} = \mathbb{1}_{(i',j') \in O_{ij}}.$$

En reprenant l'exemple de H précédent, l'orbite de la paire $(2, 3)$ est :

$$O_{23} = \{(2, 3), (3, 1), (1, 2)\} \quad \text{et} \quad M = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

Il est facile de voir que si M est la matrice d'une orbite, M' est également la matrice d'une orbite (éventuellement la même et alors $M' = M$). On définit alors l'orbite symétrisée comme $O = \dfrac{1}{2}(M + M')$.

Propriété fondamentale

Nous utiliserons régulièrement le résultat suivant : dans un groupe G les "translations" sur un groupe sont des bijections. Soit g un élément du groupe G et soit T la translation définie par

$$z \in G \to T(z) := g + z.$$

Alors cette fonction est une bijection. Quand z parcourt G , $g + z$ parcourt également G.

1.3 Modèle

On suppose, dans un modèle conceptuel (que l'on n'observe donc pas concrètement), qu'il existe une additivité entre un effet traitement et un effet unité. Si R_{ij} est la réponse conceptuelle du traitement i sur l'unité j, on pose

$$R_{ij} = \tau_i + u_j \quad \text{avec} \quad i \in \{1, \cdots, t\},\ j \in \{1, \cdots, n\},$$

où les τ_i sont les paramètres à estimer et les u_i sont des variables aléatoires dont on supposera simplement qu'elles ont un moment d'ordre 2. Pour se débarrasser d'une indétermination dans la formule ci-dessus, on suppose de plus que :

$$\sum_{i=1}^{n} \mathbb{E}\,(u_i) = 0.$$

En fait, on ne réalise pas l'allocation initiale X (matrice du plan d'expérience), mais plutôt celle obtenue après une permutation aléatoire tirée dans H. Plus précisément, si ξ est choisie au "hasard" (c'est-à-dire suivant une loi uniforme) dans H, on réalise l'allocation $\xi \cdot X$. On observe donc

$$Z = \xi \cdot X \cdot \tau + u, \quad \text{avec} \quad Z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}, \ \tau = \begin{pmatrix} \tau_1 \\ \vdots \\ \tau_t \end{pmatrix}, \ u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix},$$

et X et ξ des matrices de tailles respectives (n,t) et (n,n).

Remarque : On peut éventuellement rajouter au modèle des erreurs de mesures indépendantes. Cependant, le lecteur vérifiera qu'elles ne changent rien à l'étude suivante.

Proposition 12.1 *Avec les notations introduites précédemment, posons $Y = \xi' \cdot Z$ (où ξ' désigne la transposée de la matrice ξ, c'est-à-dire la matrice de l'application réciproque de ξ). On a alors*

$$Y = X \cdot \tau + \varepsilon \quad \text{où} \quad \varepsilon = \xi' \cdot u$$
$$\text{et} \quad \mathbb{E}(\varepsilon) = 0 \quad \text{et} \quad Var(\varepsilon) = \sum_{l=1}^{k} \gamma_l \cdot O_l,$$

où

- $O_1, \cdots, O_k$ *sont les orbites doubles symétrisées de H opérant sur Ω ;*
- $\gamma_1, \cdots, \gamma_k$ *sont des paramètres réels inconnus.*

<u>*Démonstration*</u> : On étudie l'espérance et la variance de ε_i. Comme la permutation ξ est choisie aléatoirement dans H, on a :

$$\mathbb{E}(\varepsilon_i) = \mathbb{E}\left((\xi' \cdot u)_i\right) = \frac{1}{|H|} \sum_{\xi \in H} \mathbb{E}\left(u_{\xi^{-1}(i)}\right).$$

En utilisant la propriété de groupe de H, on montre que pour tout couple (i,j) de Ω^2,

$$\#\{\xi \in H \ \text{ tel que } \ \xi(i) = j\} = \frac{|H|}{n},$$

et donc

$$\mathbb{E}(\varepsilon_i) = \frac{1}{n}\mathbb{E}(u_i) = 0. \tag{12.1}$$

Soit maintenant $\Sigma = (\sigma_{ij})_{1\leq i,j\leq n} = \mathrm{Var}\,(\varepsilon)$ la matrice de variance-covariance du vecteur ε et soit s une permutation fixe de H. Alors, pour tout (i,j) :

$$\begin{aligned}\sigma_{s(i)s(j)} &= \frac{1}{|H|}\sum_{\xi\in H}\mathbb{E}\left(u_{\xi\circ s(i)}u_{\xi\circ s(j)}\right)\\ &= \frac{1}{|H|}\sum_{\xi\in H}\mathbb{E}\left(u_{\xi(i)}u_{\xi(j)}\right) \text{ car les translations sont des bijections}\\ &= \sigma_{ij}.\end{aligned}$$

On a donc pour tout $(i,j)\in\Omega^2$, $\sigma_{s(i)s(j)} = \sigma_{ij}$ pour tout $s\in H$. Les valeurs de σ_{ij} sont donc constantes sur les orbite doubles de H ce qui implique le résultat car Σ est symétrique. ■

L'écriture du modèle obtenu dans cette proposition nous renvoie à ce que nous avons vu dans le chapitre 10 sur les modèles mixtes (Y suit présentement un modèle mixte). Avant donc d'appliquer la proposition aux plans d'expérience randomisés, revenons plus généralement aux modèles mixtes.

2 Modèles mixtes stratifiables

2.1 Présentation

Considérons le modèle mixte général $MM(X, V_1, \cdots, V_k)$ au sens du chapitre 10. On rappelle que la réponse est Y, un vecteur aléatoire gaussien de taille n, tel que

$$\begin{aligned}\mathbb{E}\,(Y) &= X\cdot\theta\\ \mathrm{Var}\,(Y) = V_\gamma &:= \sum_{i=1}^{k}\gamma_i\cdot V_i,\end{aligned}$$

où $X, V_1, \cdots, V_k$ sont des matrices connues, θ est un vecteur de p paramètres inconnus variant dans $\mathbb{R}^p$, $\gamma = (\gamma_1, \cdots, \gamma_k)$ est un vecteur de k paramètres inconnus variant dans l'ensemble $S = \{\gamma\in\mathbb{R}^k \text{ tel que } V_\gamma > 0\}$, où la dernière notation veut dire que la matrice est définie positive.

Supposons que la famille de matrices V_γ (quand γ varie dans S) admette un sous-espace propre fixe, c'est-à-dire qu'il existe E sous-espace vectoriel de $\mathbb{R}^n$ ne dépendant pas de γ tel que pour tout $x\in E$, $V_\gamma\cdot x = C_\gamma\cdot x$, avec $C_\gamma\in\mathbb{R}$.
Soit P le projecteur sur ce sous-espace fixe E. On a alors matriciellement

$$V_\gamma\cdot P = C_\gamma\cdot P.$$

Si on pose maintenant

$$\widetilde{Y} = P \cdot Y,$$

on obtient le modèle

$$\widetilde{Y} \in \mathrm{Im}(P) = E \;\text{ et }\; \mathbb{E}(\widetilde{Y}) = P \cdot X \cdot \theta, \;\; \mathrm{Var}(\widetilde{Y}) = C_\gamma \cdot P.$$

Au prix d'une paramétrisation de l'espace image $\mathrm{Im}(P) = E$, on peut montrer que ce modèle mixte peut être considéré comme un modèle linéaire ordinaire sur l'espace E dont la solution (*i.e.* estimations, tests) est très simple.

On appellera *strate* un espace propre fixe maximal (au sens de la dimension) de la famille de matrice V_γ du modèle mixte $MM(X, V_1, \cdots, V_k)$. Si les strates "recouvrent" tout l'espace au sens de la définition ci-dessous, le modèle est dit *stratifiable*. Cela amène à la propriété :

Proposition 12.2 *Soit le modèle mixte $MM(X, V_1, \cdots, V_k)$. Ce modèle est dit stratifiable si l'une des trois conditions équivalentes suivantes est vérifiée :*

i. les matrices $V_1, \cdots, V_k$ commutent ;

ii. l'espace vectoriel $[V_1, \cdots, V_k]$ engendré par les matrices $V_1, \cdots, V_k$ est une algèbre (clos par multiplication matricielle) ;

iii. pour tout $\gamma \in \mathbb{R}^k$, $V_\gamma = \sum_{i=1}^{k} \widetilde{\gamma}_i . P_i$, où les $(P_i)_{1 \le i \le k}$ forment un système orthogonal de projecteurs, c'est-à-dire que $P_i \cdot P_j = \delta_{ij} \cdot P_i$, qui vérifie $\sum_{i=1}^{k} P_i = Id$.

<u>*Démonstration*</u> : Il est immédiat de vérifier que (3) implique (1) et (2) (pour cela, on considère γ avec une seule coordonnée non nulle, et ainsi les V_i s'écrivent aussi comme des combinaisons linéaires des projecteurs P_i, qui commutent et forment une algèbre). Pour démontrer que (2) implique (1), il suffit de remarquer que le produit de matrices $V_i.V_j$ est forcément une matrice symétrique et cela implique que les matrices V_i commutent (car elles sont elles-mêmes symétriques).
Pour démontrer que (1) implique (3), on remarque que comme les matrices $V_1, \cdots, V_k$ sont symétriques, elles sont diagonalisables, et comme elles commutent, elles sont diagonalisables dans une même base. On considère ensuite la partition la moins fine des valeurs propre telle qu'à l'intérieur de chaque sous-espaces de cette partition les valeurs propres sont toutes égales pour chacune des matrices V_i. On définit les projecteurs P_i comme étant les projecteurs orthogonaux sur les sous-espaces de la partition. Soit m le nombre de ces projecteurs. Par un argument de dimension, $m = k$.
■

2.2 Analyse en strates d'un modèle mixte stratifiable

Soit un modèle mixte stratifiable. Alors l'application qui aux données $Y \in \mathbb{R}^n$ associe les k vecteurs de $\mathbb{R}^n$, $P_1 \cdot Y, \cdots, P_k \cdot Y$ est une bijection. De plus, les $P_i \cdot Y$ pour $i = 1, \cdots, k$ suivent des modèles linéaires ordinaires non corrélés, car on a vu alors que $\mathbb{E}(P_i \cdot Y) = P_i \cdot X \cdot \theta$ et $\mathrm{Var}(P_i \cdot Y) = C_\gamma \cdot P_i$ et $\mathbb{E}(P_i \cdot Y \cdot (P_j \cdot Y)') = 0$ dès que $i \neq j$.

Pour chaque $i \in \{1, \cdots, k\}$, on obtient dans la strate i (donc en opérant la projection P_i) un estimateur de θ (ou d'une partie de θ car certaines parties peuvent ne pas être estimables). On peut montrer que si γ est connue, l'estimateur optimal des moindres carrés généralisés (voir chapitre 10) est une combinaison des estimateurs de θ dans chaque strate. Cette combinaison peut être très complexe dans certains cas et très simple dans d'autres, comme on le verra dans les exemples qui vont suivre.

2.3 Cas du modèle randomisé

Nous considérons à nouveau le modèle issu de la proposition 12.1. Ce modèle est un modèle mixte que l'on peut noter $MM(Y, O_1, \cdots, O_k)$. Les matrices $O_1, \cdots, O_k$ sont invariantes par H, c'est-à-dire que si ξ_0 est une permutation quelconque de H, alors $\xi_0 \cdot O_i \cdot \xi_0' = O_i$.
On pourrait montrer, mais nous ne le détaillerons pas ici, que travailler sur les données prises dans un ordre aléatoire Y telles que nous les avons considérées ou dans l'ordre réel Z est équivalent tant que l'on se limite à l'analyse d'un modèle mixte telle qu'elle a été définie au chapitre 10 . En conclusion, tout se passe comme si les données prises dans l'ordre réel Z suivaient le même modèle mixte.

Pour savoir si le modèle randomisé est stratifiable et pour trouver sa décomposition, on utilisera le plus souvent le critère suivant :

Proposition 12.3 *Supposons que l'on ait déterminé k projecteurs orthogonaux entre eux $P_1, \cdots, P_k$ dont la somme vaut l'identité (au sens de la condition (3) de la définition 12.2). Si les espaces $Im(P_1), \cdots, Im(P_k)$ sont invariants par H, alors le modèle est stratifiable et, au prix d'un éventuel regroupement, les $P_1, \cdots, P_k$ sont les projecteurs sur les strates.*

La démonstration, qui ne sera pas détaillée, réside dans le fait que tout projecteur sur un espace invariant commute avec les éléments de H.

Nous allons maintenant appliquer ces résultats aux plans d'expériences randomisés classiques.

3 Application aux plans d'expériences randomisés

3.1 Plan en randomisation totale

Dans ce plan, nous avons (avec les notation du chapitre précédent 11) $n = r \cdot t$. Le plan peut être défini à partir d'une allocation initiale triviale où les r premières unités reçoivent le traitement 1, ensuite, les r suivantes le traitement 2, etc... En conséquence, la matrice X s'écrit :

$$X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

D'après la description du plan en randomisation totale, le groupe H est alors le groupe symétrique des $n!$ permutations. Il est clair que pour ce groupe de permutations, il n'y a que deux orbites doubles :

- la première orbite est constituée des couples (i, i) avec $i \in \Omega$ et a pour matrice la matrice identité Id ;
- la seconde orbite comprend les couples (i, j) avec $(i, j) \in \Omega^2$ et $i \neq j$ et a donc pour matrice une matrice avec que des 1 sauf sur la diagonale sur laquelle il y a des 0. Cette matrice est parfois note $J - Id$.

Notons que ces orbites doubles symétriques. D'après la proposition 12.2(iii), la matrice de variance-covariance du modèle peut s'écrire

$$V_\gamma = \widetilde{\gamma}_1 P_{\text{II}} + \widetilde{\gamma}_2 (Id - P_{\text{II}}),$$

où $\widetilde{\gamma}_1$ et $\widetilde{\gamma}_2$ sont deux paramètres inconnus.

On applique la proposition 12.3 aux deux projecteurs suivants :

- le projecteur sur la moyenne générale, P_{II}, qui à $(Y_{ij})_{ij}$ associe $(\overline{Y})_{ij} := (Y_{..})_{ij}$;

– et le projecteur $Id - P_{\text{II}}$.

Ces deux projecteurs sont les projecteurs sur les deux strates. On applique maintenant deux considérations qui sont relativement générales :

i. la projection sur la premiere strate donne le modèle :

$$(P_{\text{II}} \cdot Y)_i = (Y_\cdot)_i = (P_{\text{II}} \cdot X \cdot \theta)_i + \varepsilon = (t_\cdot) + \varepsilon,$$

où $t_\cdot$ est la moyenne des traitements. On considère que cette projection est sans intérêt car elle ne permet pas de faire des comparaisons entre les traitements, donc on l'oublie purement et simplement.

ii. pour la seconde strate, on utilise la notion de modèle équivalent : on numérote les données par un premier indice qui est celui du traitement et par un second indice celui de répétition, et on considère le modèle

$$Z_{ij} = t_i + e_{ij}, \tag{12.2}$$

où les e_{ij} sont centrés, non corrélés et de même variance. Le lecteur peut vérifier que ce modèle a exactement la même projection sur la seconde strate que le modèle randomisé (c'est-à-dire le même modèle mais avec les variables recentrées). Comme la première strate est considérée comme vide, nous arrivons à la conclusion que le modèle randomisé est équivalent au modèle 12.2.

3.2 Plan en blocs complets

Nous considérons maintenant ce type de plan, dont l'étude théorique est plus complexe que celle du plan précédent. La longueur des calculs fait que tous les détails ne seront pas donnés.

Pour un plan en blocs complets, le nombre d'unités est encore $b \cdot k = r \cdot t$. On part d'une allocation initiale triviale, par exemple, on place tous les traitements dans le même ordre $1, 2, 3, \cdots$ dans chaque bloc. Comme dans un plan en blocs complets, il y a randomisation des blocs et des unités dans les blocs, le groupe H est maintenant ce que les algébristes appellent le produit en couronne du groupe symétrique S_b par le groupe symétrique S_k. Plus simplement, un élément de ce groupe peut être décrit comme une certaine permutation des b blocs composée avec b permutations intra-blocs de taille k. On montre qu'il y a trois orbites doubles :

– la première orbite est constituée par toutes les couples d'unités identiques, soit l'ensemble $\{(i, i),\ 1 \leq i \leq n\}$;

– la seconde orbite comprend les couples d'unités différentes prises dans le même bloc, soit l'ensemble $\{(i,j) \in b_k^2,\ i \neq j,\ 1 \leq k \leq b\}$;

– la troisième orbite comprend les couples d'unités prises dans deux blocs différents, soit l'ensemble $\{(i,j) \in (b_k \times b_{k'}),\ 1 \leq k, k' \leq b,\ k \neq k'\}$.

On identifie les trois strates associées :

– à la projection sur la moyenne générale, soit P_{II} ;

– à la moyenne par bloc orthogonalement à la moyenne générale, soit $P_b - P_{\mathrm{II}}$, où P_b est la projection qui à chaque élément d'un bloc associe la moyenne du bloc ;

– à la projection dite intra-bloc, c'est-à-dire, le recentrage des données par bloc, donc $Id - P_b$ (orthogonalement aux deux projections précédentes).

On montre que les deux premières strates sont sans intérêt (elles ne permettent pas de différencier les traitements) et que la troisième est équivalente au modèle additif traitement-bloc de la partie 3.2.

3.3 Plan en blocs incomplets

Ici, on remarque que le groupe H est le même que précédemment, seule change l'allocation initiale (la matrice X). En conséquence, les strates sont les mêmes. On en déduit deux approches du plan en blocs incomplets :

i. l'approche moderne qui consiste à remarquer que la projection sur la première strate reste sans intérêt, et que sur les deux autres strates, le modèle est bien équivalent au modèle à effets bloc aléatoires aléatoire de la partie 3.3 du chapitre 11. On valide donc les affirmations de cette partie. On peut donc traiter les résultats de l'expérience par des méthodes de modèles mixtes. La décomposition en strates ne sert qu'à valider le modèle.

ii. l'approche classique, qui, dans le cas de blocs incomplets et a contrario du cas de blocs complets, est confronté au fait que la seconde strate contient des informations sur les comparaisons entre traitements. On décompose alors le modèle en deux parties : 1/ On montre que les projections du modèle randomisé sur les deux première strates sont équivalentes à un modèle à erreurs indépendantes sur les sommes par blocs. Ce modèle est appelé par abus de language *la strate inter-bloc* dans la mesure où pour l'école anglaise il est évident que la première strate est sans intérêt ; 2/ On montre que la projection sur la troisième strate,

dite *strate intra-bloc*, est équivalente au modèle avec effets traitements et blocs fixes. En effet nous laissons le soin au lecteur de vérifier que les projections sur les deux premières strates de ce dernier modèle ne contiennent aucune information sur les traitements. Quant à la projection sur la troisième strate, elle est clairement équivalente pour le modèles à effets bloc fixe et pour modèle mixte issu de la randomisation.

4 Exercices

Exercice 12.1

(***) [Plans en ligne et colonnes] Dans ce plan, les unités sont rangées dans un réseau de r lignes et de c colonnes. Un élément de H correspond à la composition d'une permutation des ligne et d'une permutation des colonnes.

i. Montrer que H est simplement transitif.

ii. Montrer qu'il y a 4 orbites doubles, que le modèle est stratifiable et identifier les 4 strates.

iii. On suppose que le plan est un carré latin. Montrer que l'information est totalement contenue dans la dernière strate et que le modèle est bien équivalent à un modèle additif à trois facteurs : traitement, ligne et colonne.

iv. On suppose que le plan n'est plus complet en lignes et en colonnes. Montrer l'équivalence avec le modèle mixte à lignes et colonnes aléatoires.

Exercice 12.2

(***) [Plan split-plot] Ici, les "sous-parcelles" sont regroupées s par s en "grandes parcelles" qui sont elles-mêmes regroupées t par t en blocs. Au final on dispose de r blocs. Un élément de H correspond à un mélange des r blocs, composé à r permutations différentes des grandes parcelles intra-bloc, composé encore avec $r \cdot t$ permutations des sous-parcelles à l'intérieur des grandes parcelles.

i. Montrer que H est simplement transitif, qu'il y a 4 orbites doubles et donc 4 strates.

ii. Montrer l'équivalence avec le modèle donné au chapitre 11.

iii. Montrer que le second facteur et l'interaction sont totalement estimables dans la strate dite "intra-grande parcelle".

iv. Montrer que le premier facteur est totalement estimable dans la strate dite "inter-grandes parcelles".

Pour plus de détails sur ce sujet on pourra consulter les articles de Bardin et Azaïs [9], ou de Bailey [6].

Chapitre 13

Plans fractionnaires

Nous présentons une méthode pour étudier un phénomène quantitatif influencé par un certain nombre p de variables qualitatives ou facteurs. Nous définissons les notions d'interactions dans un cadre général, puis, dans le cas où les p facteurs ne prennent que deux valeurs, nous montrons comment construire un plan factoriel fractionnaire qui consiste à réaliser un sous-ensemble, une fraction, de l'ensemble complet, de taille 2^p, de toutes les combinaisons entre les facteurs.

1 Introduction

Lors du chapitre 11, nous avons décrit des situations dans lesquelles la préparation des données par un plan randomisé permettait d'améliorer l'étude des résultats des expériences. Cependant, les techniques que nous avons étudiées concernaient un, deux, voire au maximum trois facteurs. Nous allons maintenant proposer une autre méthode de planification des expériences permettant de prendre en compte un nombre quelconque de facteurs.

Exemple : on considère une réaction chimique qui dépend pour simplifier de trois facteurs : le ph PH, avec une valeur standard de 7, la température, T, avec comme valeur standard $30°C$, et la dose D, avec comme valeur standard 100. On sait que l'on peut faire varier chacun de ces facteurs entre deux limites et on cherche à savoir s'ils ont une influence sur la réponse, par exemple le rendement de la réaction chimique. On va comparer deux expériences :

- Expérience 1 : On fait varier d'abord le facteur PH. On réalise 4 répétitions à $PH = 6.5$, $T = 30°C$, $D = 100$ et 4 répétitions à $PH = 7.5$, $T = 30°C$,

$D = 100$.
On fait varier ensuite le facteur température. On effectue 4 répétitions à $PH = 7$, $T = 25°C$, $D = 100$ et 4 répétitions à $PH = 7$, $T = 35°C$, $D = 100$.
Enfin, on fait varier le facteur dose. On effectue 4 répétitions à $PH = 7$, $T = 30°C$, $D = 90$ et 4 répétitions à $PH = 7$, $T = 30°C$, $D = 110$.

Le coût total de cette expérience est de 24 unités et la "puissance expérimentale" [1] est la comparaison de moyennes de 4 éléments : pour tester la significativité d'un facteur, on compare dans la sous-expérience correspondante les 4 données reliées au niveau haut avec les 4 données reliées au niveau bas.

- Expérience 2 : on effectue les 8 combinaisons possibles entre les deux valeurs (hautes et basses) de chacun des trois facteurs. C'est-à-dire les huit unités :

$$\begin{array}{cccc} (6.5, 25°C, 90) & (6.5, 25°C, 110) & (6.5, 35°C, 90) & (6.5, 35°C, 110) \\ (7.5, 25°C, 90) & (7.5, 25°C, 110) & (7.5, 35°C, 90) & (7.5, 35°C, 110) \end{array}$$

Le coût total est maintenant de 8 unités. Si on fait l'analyse à l'aide d'un modèle additif à trois facteurs : H, T et D, ce modèle est orthogonal. On comparera donc des moyennes de 4 observations pour tester la significativité d'un facteur. Donc en première approximation, on a construit une expérience trois fois moins chère (en terme de nombres de mesures à effectuer) que l'expérience 1 et qui a pourtant la même "puissance expérimentale".

Remarquons bien qu'en présence d'interactions entre les facteurs, il sera possible de les détecter dans la seconde expérience : en effet, il reste 4 degrés de liberté dans la somme des carrés résiduelle, il est donc encore possible de consacrer un ou deux degrés de liberté pour un ou deux termes d'interaction. En revanche, comme l'expérience 1 est subdivisée en sous-expériences dans lesquelles on ne fait varier qu'un facteur à la fois, il ne sera pas possible de détecter la présence d'interaction.

En conclusion de cet exemple, lorsque un phénomène dépend de plusieurs facteurs, on a toujours intérêt à les étudier globalement. Le plan le plus simple est celui de l'expérience 2 où l'on fait toutes les combinaisons des niveaux (dans notre exemple chaque facteur a deux niveaux) de tous les facteurs. Il est appelé *plan factoriel complet*.

S'il y a p facteurs possédant tous 2 niveaux, le plan factoriel complet demande 2^p unités. Cela est souvent trop coûteux. On ne va donc réaliser qu'une partie, une fraction, du plan complet. Un tel plan est alors appelé *plan factoriel fractionnaire*.

1. Nous ne voulons pas donner de définition précise de cette notion. Tout ce dont nous avons besoin et de savoir que plus on fait la moyenne d'un grand nombre de répétitions, plus précise est l'estimation

2 Cadre général pour des facteurs à deux niveaux

On considère p facteurs à deux niveaux, que l'on notera -1 et $+1$ sans perte de généralité. On appelle traitement une combinaison $(i_1, \cdots, i_p)$ de ces facteurs. L'ensemble des traitements est de cardinal $n := 2^p$ et correspond au plan factoriel complet. Soit E l'espace des réponses aux divers traitements. On a ainsi

$$E = \{f,\ f : \{-1,+1\}^p \mapsto \mathbb{R}\} \quad \underset{\text{(isomorphe)}}{\longleftrightarrow} \quad \mathbb{R}^{2^p}$$

(on vérifiera aisément que ces deux ensembles sont des espaces vectoriels en bijection). Pour deux fonctions f et g de E, on définira la norme $||f||$ et le produit scalaire $< f, g >$) à partir de la norme et du produit scalaire euclidien de l' écriture dans $\mathbb{R}^{2^p}$. Par exemple, pour $p = 2$, et pour f telle que $f(1,1) = -3.2$, $f(1,-1) = -14.72$, $f(-1,1) = -2$ et $f(-1,-1) = \pi$, on a : $||f||^2 = (-3.2)^2 + (-14.72)^2 + (-2)^2 + \pi^2$.

On se place dans le cas où le nombre total ($n = 2^p$) des expériences à réaliser pour obtenir un plan factoriel complet est trop important pour que le plan puisse être mis en place (par exemple pour des raisons de coût de chaque expérience). On va n'effectuer qu'une partie, une fraction de ce plan, ce qui conduira à la mise en place d'un *plan factoriel fractionnaire*.

En premier lieu, nous allons définir proprement les interactions multiples et les effets principaux. Dans E, il existe des éléments particuliers : les fonctions coordonnées. On définit la k-ème fonction coordonnée, notée A_k, pour $k = 1, \cdots, p$, par :

$$A_k(i_1, \ldots, i_p) = i_k; \quad (i_1, \ldots, i_p) \in \{-1, 1\}^p.$$

(cela définit clairement une fonction de E). On obtient facilement la propriété suivante :

$$< A_i, A_j >= \begin{cases} n := 2^p & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

(dans le cas où $i \neq j$, on vérifie qu'il y a $n/4$ p-uplets de $\{-1,+1\}^p$ tels tels que $A_i \cdot A_j = (1,1)$, puis $n/4$ p-uplets de $\{-1,+1\}^p$ tels tels que $A_i \cdot A_j = (1,-1)$, puis $n/4$ tels que $A_i \cdot A_j = (-1,1)$ et enfin $n/4$ tels que $A_i \cdot A_j = (-1,-1)$; Le produit scalaire vaut bien 0).
Soit $B \subset \{1, \cdots, p\}$. On définit la fonction A^B, fonction appartenant également à E et telle que pour $x \in \{-1,1\}^p$,

$$A^B(x) := \prod_{i \in B} A_i(x) = A_1^{\varepsilon_1}(x) \times \cdots \times A_p^{\varepsilon_p}(x) \quad \text{où } \varepsilon_k = 1 \text{ lorsque } k \in B \text{ et } \varepsilon_k = 0 \text{ sinon}$$

(par convention, la fonction A_k^0 est la fonction constante égale à 1. telle que $\forall x \in \mathbb{R}^p$, $A_k^0(x) = 1$). On a alors la proposition suivante :

Proposition 13.1 *Quand B varie dans l'ensemble des parties de $\{1,\ldots,p\}$, les A^B forment une base orthogonale de E de norme $\sqrt{n} = \sqrt{2^p}$.*

Démonstration : Soit $B \subset \{1,\ldots,p\}$ il est clair que A^B vu comme un élément de $\mathbb{R}^n$ a toutes ses composantes qui valent ± 1 et donc le carré de sa norme vaut n. D'autre part, soit B' dans $\{1,\ldots,p\}$ distinct de B. Alors, il existe un élément $j \in \{1,\ldots,p\}$ qui appartient à l'un de ces deux ensembles et pas à l'autre. Quitte à renuméroter les facteurs, on peut toujours supposer que $j = p$. On peut écrire que :

$$A^B(x) = A_1^{\varepsilon_1}(x) \times \cdots \times A_p^{\varepsilon_p}(x).$$

Posons maintenant :

$$A^{B'}(x) = A_1^{\varepsilon'_1}(x) \times \cdots \times A_p^{\varepsilon'_p}(x).$$

On a alors $\varepsilon_p + \varepsilon'_p = 1$ et

$$< A^B, A^{B'} > = \sum_{i_1,\ldots,i_p=\pm 1} i_1^{\varepsilon_1+\varepsilon'_1} \ldots i_p^{\varepsilon_p+\varepsilon'_p} = \sum_{i_1,\ldots,i_{p-1}=\pm 1} i_1^{\varepsilon_1+\varepsilon'_1} \ldots i_{p-1}^{\varepsilon_{p-1}+\varepsilon'_{p-1}}((+1)+(-1)) = 0,$$

où $((+1)+(-1))$ est la somme pour $i_p = \pm 1$ de $i_p^{\varepsilon_p+\varepsilon'_p}$. ■

Soit B un sous-ensemble de $\{1,\cdots,p\}$. On définit maintenant les sous-espaces vectoriels de E suivants :

- V_B est l'ensemble des fonctions de $\{-1,1\}^p$ dans $\mathbb{R}$ ne dépendant que des seules coordonnées présentes dans B. Si $B = \{k_1,\cdots,k_m\}$, une fonction f de V_B s'écrit $f(x) = c(x_{k_1},\cdots,x_{k_m})$;

- W_B est le sous-espace vectoriel de V_B orthogonal à tous les sous-espaces V_D quand $D \subset B,\ D \neq B$.

Par exemple, si $p = 2$, alors :

- $E = \{f : \{-1,1\}^2 \mapsto \mathbb{R}\}$, donc $f \in E$ est telle qu'il existe $(a,b,c,d) \in \mathbb{R}^4$ tel que $f(1,1) = a$, $f(-1,1) = b$, $f(1,-1) = c$ et $f(-1,-1) = d$.

- une fonction de $V_{\{1\}}$ est telle qu'il existe $(a,b) \in \mathbb{R}^2$ vérifiant $f(1,1) = f(1,-1) = a$ et $f(-1,1) = f(-1,-1) = b$.

- une fonction de $W_{\{1\}}$ est telle qu'il existe $a \in \mathbb{R}$ vérifiant $f(1,1) = f(1,-1) = a$ et $f(-1,1) = f(-1,-1) = -a$ (le seul sous-ensemble strictement inclus dans $\{1\}$ est $\emptyset$, et $V_\emptyset = \{f \in E, f = a$ avec $a \in \mathbb{R}\}$, fonctions constantes).

On a alors les propriétés suivantes quant à ces sous-espaces vectoriels de E :

Proposition 13.2 *Soit B un sous-ensemble de $\{1, \cdots, p\}$. Alors :*

i. $dim(V_B) = 2^{|B|}$, *où $|B|$ est le cardinal de B ;*

ii. si $D \subset B$ alors $A^D \in V_B$;

iii. l'ensemble $\left\{A^D \text{ avec } D \subset B\right\}$ engendre V_B ;

iv. $A^B \in W_B$ et A^B engendre W_B qui est donc de dimension 1.

<u>*Démonstration*</u> : les propositions i et ii découlent de l'écriture même d'une fonction de V_B. La proposition iii découle de i et ii et de l'orthogonalité des A^D où $D \subset B$. Enfin, comme on a $A^B \in V_B$ et pour D strictement inclus dans B, A^B et A^D sont orthogonales alors $A^B \in W_B$. Puisque A^B et les A^D forment une base de V_D (proposition (iii)), A^B engendre bien W_B. ■

Définition 13.1 *L'espace W_B sera appelé espace de l'interaction entre les facteurs de l'ensemble B en adoptant les conventions suivantes : si B ne contient qu'un facteur, l'espace correspondant, que l'on appellera interaction d'ordre* 1, *est l'espace associé à l'effet principal de ce facteur ; $W_\emptyset$ est l'espace de la "moyenne générale" considéré comme une interaction d'ordre* 0.

W_B est donc de dimension 1 quelle que soit la taille de B.

Définition 13.2 *La vraie réponse f (a priori non connue) au traitement est un élément de E qui se décompose sur la base des générateurs A^B. Nous appellerons valeur de l'effet de A^B (ou valeur d'un effet principal ou d'une interaction entre les facteurs de B), notée $e(A^B)$, le coefficient de cette décomposition et on a :*

$$f = \sum_{B \subset \{1, \cdots, p\}} e(A^B) A^B.$$

Le terme "interaction" sera en fait un terme générique. Nous avons défini l'espace de l'interaction triple par exemple entre les facteurs A_1, A_2, A_3 nous avons défini la valeur de l'effet de l'interaction. Le terme "interaction" tout court fait référence à l'un ou l'autre suivant le contexte.

Exemple avec trois facteurs : Le tableau suivant décrit le plan factoriel complet ainsi que les différents générateurs avec nos notations :

$A^{\emptyset}$	$A^{\{1\}}$	$A^{\{2\}}$	$A^{\{3\}}$	$A^{\{1,2\}}$	$A^{\{1,3\}}$	$A^{\{2,3\}}$	$A^{\{1,2,3\}}$	traitement
1	-1	-1	-1	$+1$	$+1$	$+1$	-1	$(-1,-1,-1)$
1	-1	-1	$+1$	$+1$	-1	-1	$+1$	$(-1,-1,+1)$
1	-1	$+1$	-1	-1	$+1$	-1	$+1$	$(-1,+1,-1)$
1	-1	$+1$	$+1$	-1	-1	$+1$	-1	$(-1,+1,+1)$
1	$+1$	-1	-1	-1	-1	$+1$	$+1$	$(+1,-1,-1)$
1	$+1$	-1	$+1$	-1	$+1$	-1	-1	$(+1,-1,+1)$
1	$+1$	$+1$	-1	$+1$	-1	-1	-1	$(+1,+1,-1)$
1	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$+1$	$(+1,+1,+1)$

Notons $A := A^{\{1\}},\ B := A^{\{2\}},\ C\ := A^{\{3\}}$.
Maintenant pour des raisons de budget nous décidons de ne conserver que 4 unités. Choisissons par exemple les 4 unités pour lesquelles $A \cdot B \cdot C = 1$, ce qui définit un "demi plan" ou "fraction". Même en supposant que l'on observe la réponse sans erreur, on est amené à résoudre le système de quatre équations :

$$
\begin{aligned}
f(-1,-1,+1) &= e(1) + e(A\cdot B\cdot C) - e(A) - e(B\cdot C) - e(B) - e(A\cdot C) + e(C) + e(A\cdot B)\\
f(-1,+1,-1) &= e(1) + e(A\cdot B\cdot C) - e(A) - e(B\cdot C) + e(B) + e(A\cdot C) - e(C) - e(A\cdot B)\\
f(+1,-1,-1) &= e(1) + e(A\cdot B\cdot C) + e(A) + e(B\cdot C) - e(B) - e(A\cdot C) - e(C) - e(A\cdot B)\\
f(+1,+1,+1) &= e(1) + e(A\cdot B\cdot C) + e(A) + e(B\cdot C) + e(B) + e(A\cdot C) + e(C) + e(A\cdot B)
\end{aligned}
$$

D'après les relations d'orthogonalité déjà prouvées, ce système a une solution unique pour les nouvelles inconnues :

$$
\left\{
\begin{array}{lcl}
x_1 &=& e(1) + e(A\cdot B\cdot C)\\
x_2 &=& e(A) + e(B\cdot C)\\
x_3 &=& e(B) + e(A\cdot C)\\
x_4 &=& e(C) + e(A\cdot B)
\end{array}
\right. .
$$

On a regroupé les effets ayant le même signe.

En conclusion, cet exemple montre que dans le tableau restreint correspondant à $A \cdot B \cdot C = 1$, certaines relations sont vérifiées entre les vecteurs générateurs. Elles découlent directement de la relation qui a défini la fraction : $1 = A \cdot B \cdot C,\ A = B \cdot C,\ B = A \cdot C,\ C = A \cdot B$. Elles impliquent directement les confusions observées entre les différentes interactions. En effet, dans l'écriture générale :

$$
f = \sum_{B \subset \{1,\cdots,p\}} e(A^B) A^B,
$$

certains (A^B) sont confondus sur la fraction choisie et on regroupe donc les coefficients $e(A^B)$ correspondants.

3 Méthode des facteurs de base

On appellera plan $2^{p-q}(p > q)$, un plan comprenant 2^{p-q} unités pour p facteurs à deux niveaux. Le taux de fraction est donc 2^{-q}. Une méthode de construction qui donne des fractions régulières (terme que l'on définira ultérieurement) est la méthode des facteurs de base : on construit le plan factoriel complet pour les $(p-q)$ premiers facteurs. On définit ensuite les valeurs des autres facteurs en fonction des $p-q$ premiers (sous forme de produit exclusivement). On appelle *clef du plan* l'ensemble de ces q relations que l'on exprimera sous la forme canonique $1 = ...$ Par exemple, $F = A \cdot B$ sera écrit $1 = A \cdot B \cdot F$. Il est facile de vérifier que la fraction ainsi définie est exactement le sous-ensemble du plan complet qui vérifie la clef. On définit également les *relations complètes de définition* qui sont les 2^q terme égaux que l'on obtient en combinant les relations issues de la clef. Ces relations donnent l'*alias* de 1, c'est-à-dire les termes confondus avec 1. A partir de cet alias et par combinaison, on obtient l'alias d'un terme quelconque donné qui comprend $2^q - 1$ termes.

Exemple 13.1 *Considérons le plan 2^{5-2} pour 5 facteurs A, B, C, D et E avec A, B, C comme facteurs de base et la clef : $D = A \cdot B$ et $E = A \cdot C$. Il s'agit donc d'un quart de plan à 8 unités. Les relations complètes de définition sont*

$$1 = A \cdot D \cdot B = A \cdot C \cdot E = D \cdot C \cdot B \cdot E.$$

En effet, $A^2 = B^2 = \cdots = E^2 = 1$. Il y a bien 4 effets confondus. L'alias de A par exemple comprend trois termes :

$$A = D \cdot B = C \cdot E = A \cdot B \cdot C \cdot D \cdot E.$$

Il y a donc, en particulier, confusion dans cette fraction entre l'effet principal du facteur A et les interactions $D * B$ et $C * E$.

Estimation des effets : On suppose que les observations sont faites avec une erreur de variance σ^2 et que les erreurs ne sont pas corrélées. Choisissons une suite de générateurs :$A^{B_1}, \cdots, A^{B_m}$ avec $m < 2^{p-q}$, telle qu'il y ait au plus un représentant de chaque alias (il ne faut pas les prendre tous si on veut que le modèle linéaire ne soit pas saturé). Considérons le modèle linéaire

$$Y(i_1, \cdots, i_p) = \sum_{i=1}^{m} \widetilde{e}(A^{B_i}) A^{B_i} + \varepsilon_{i_1, \cdots, i_p},$$

où $\widetilde{e}(A^{B_i})$ est l'alias de $e(A^{B_i})$ au sens de la somme de effets de tous les interactions qui lui sont confondues et où $(i_1, \cdots, i_p)$ varie sur la fraction.

Alors ce modèle est orthogonal sur l'ensemble des données du plan fractionnaire. La matrice d'information : $X' \cdot X$ vaut $n \cdot Id$ (n est le nombre de données, soit $n = 2^{p-q}$) et les estimateurs

$$\widehat{\widetilde{e}}(A^{B_i}) = \frac{< Y, A^{B_i} >}{n},$$

sont non corrélés et de variance σ^2/n. Notez que le produit scalaire est dans $\mathbb{R}^{2^{p-q}}$ dans la formule ci-dessus

Une fraction ayant la propriété de n'avoir que des effets orthogonaux ou confondus est appelée une *fraction régulière* . Toutes les fractions présentées dans ce chapitre sont régulières à l'exception de celle de l'exercice 6.

4 Plan pour l'étude des effets principaux et des interactions doubles

Définition 13.3 *Un plan est dit de résolution ρ entière lorsque*

$$\rho = \inf \{\textit{nombre de symboles des éléments de l'alias de } 1\}.$$

La résolution est notée traditionnellement en chiffre romain.

Exemples : Résolution III : tous les effets principaux sont non confondus.
Résolution IV : un effet principal ne peut être confondu avec une interaction, mais deux interactions peuvent être confondues.
Résolution V : on peut poser un modèle avec toutes les interactions et les effets principaux sans confusion.
Nous laissons au lecteur le soin de vérifier l'exactitude de ces affirmations. De manière générale, la résolution V est considérée comme suffisante dans toutes les situations, alors que la résolution III est considérée comme une propriété minimale.

Définition 13.4 *L'aberration d'un plan pour p facteurs est un p-uplet dont l'élément i contient le nombre de symboles de i lettres dans l'alias de 1 (1 exclu).*

Par exemple l'aberration du plan de l'exemple 13.1 vaut : $(0, 0, 2, 1, 0)$. Cette notion d'aberration est un raffinement de la notion de résolution. Les aberrations sont munies de l'ordre lexicographique et on parle ainsi de "plan de minimum d'aberration". Par exemple il est facile de vérifier que l'aberration du plan 2^{7-2} de la table 4 vaut (0001200). Si on considère le plan 2^{7-2} donné par les clefs $F = ABC$, $G = ABD$, il est encore de résolution IV mais son aberration vaut maintenant (0003000) ce qui est moins bien.

Proposition 13.3 • *Si le plan est de résolution III, tous les effets principaux sont estimables ; en comptant les degrés de liberté(la dimension du modèle linéaire) le nombre p de facteurs vérifie $p \leq n-1$. Ce maximum est atteint dans le sens où pour tout $r \in \mathbb{N}^*$, il existe un plan de résolution III avec 2^r unités et $2^r - 1$ facteurs.*

• *Si le plan est de résolution IV avec 2^r unités, alors il comprend au maximum 2^{r-1} facteurs. Ce maximum est atteint.*

• *Pour un plan de résolution V, il n'y a pas de résultat général. On peut simplement donner le nombre de facteurs maximaux pour les petites valeurs de r :*

nombre r	4	5	6	7	8	9
nombres d'unités 2^r	16	32	64	128	256	512
nb max. de facteurs p	5	6	8	11	17	23
d.d.l. du modèle	16	22	37	67	154	277
d.d.l. de (modèle $+1$ *facteur)*	22	29	46	79	172	301

L'avant dernière ligne donne la dimension du modèle complet avec toutes les interaction d'ordre 2. La dernière ligne donne la dimension du modèle avec un facteur de plus. On voit donc que le seul critère de dimension montre que la solution proposée est maximale uniquement dans le cas de 16 unités.

Démonstration : La réciproque du premier point est facile : à partir des r facteurs de base, on définit tous les facteurs complémentaires a partir de tous les produits de deux ou plus facteurs de base. Nous démontrons le second point. Soit

$$D := \{\text{produit d'un nombre impair de facteurs parmi } r\}.$$

Alors,

(i) le produit de 3 éléments de D ne vaut jamais 1 (raisonner par parité) ;

(ii) le cardinal de D est 2^{r-1} a) si r est impair c'est évident : à tout produit impair correspond le produit complémentaire qui est pair, b) si r est pair : on isole un des facteurs et on se ramène au cas a) ;

(iii) on ne peut pas trouver un ensemble plus grand que D dans le sens où si $C_1, \ldots, C_k$ est un ensemble de k générateurs vérifiant (i) alors $C_1, \ldots, C_k, C_1^2, \ldots, C_1 \ldots C_k$ (par exemple) sont tous différents et donc $k \leq 2^{r-1}$.

Notez que D comprend tous le singletons. On définit donc r facteurs de base et $2^{r-1} - r$ facteurs complémentaires à l'aide des clefs qui correspondent aux éléments de D qui ne sont pas des singletons. La propriété (iii) montre que l'on ne peut pas faire mieux. ■

Compléments :

- Quand le plan est de taille trop importante pour pouvoir être conduit de manière homogène, il doit être découpé en blocs . Les facteurs blocs peuvent être considérés comme des facteurs ordinaires sauf que les notions de résolution et d'aberration n'ont plus la même pertinence. Le mieux est de regarder en détail les confusions, en étant bien conscient que tout effet confondu avec un facteur bloc sera totalement non-estimable intra-bloc (voir exemple ci-dessous).

- Si un ou plusieurs des facteurs ont un nombre de niveaux égal à 4, 8 (ou plus généralement une puissance de 2), on peut se ramener au cas précédent en recodant les niveaux à l'aide de 2 ou 3 pseudo-facteurs : par exemple si A possède 4 niveaux notés $1, 2, 3, 4$, on peut le recoder à l'aide de 2 pseudo-facteurs en utilisant la table suivante :

A	A_1	A_2
1	1	1
2	1	-1
3	-1	1
4	-1	-1

- Dans le cas où les facteurs ont plus de deux niveaux, les plans fractionnaires utilisent dans le cas général des techniques de corps finis qui sont basées sur des nombres premiers. Pour ces raisons, il n'est possible que de travailler que sur un seul nombre premier : on peut donc faire des plans pour des facteurs à 2, 4 ou 8 niveaux comme pour des facteurs à 3, 9 ou 27 . niveaux, mais il est impossible de mélanger les deux. Pour des facteurs à 3 niveaux, il faut travailler non plus avec $\{-1, 1\}$ mais avec les racines cubiques de l'unité $\{1, j, j^2\}$, ce qui amène à une présentation plus technique.

- **Méthode de Tagushi** Certains facteurs d'une expérience que l'on peut contrôler en laboratoire peuvent ne pas être contrôlés en utilisation normale. Un des apports de Tagushi est d'avoir proposé des plans pour étudier l'influence de ces facteurs sur la variabilité du résultat. Par exemple, dans un atelier de construction, on veut régler une fraise pour une certaine performance donnée. Mais dans l'utilisation future de cette fraise il y aura certains facteurs incontrôlés : la température de l'atelier, le degré d'usure de la fraise, la température de l'huile de refroidissement.
  On construit ainsi un plan fractionnaire pour les facteurs contrôlés en utilisation normale et un autre plan fractionnaire pour les facteurs non contrôlés en utilisation normale. On "croise" ensuite les plans. Si le dispositif doit être réglé sur une spécification précise, on va rechercher le réglage de facteurs contrôlés en utilisation normale qui minimise l'Erreur Quadratique Moyenne, c'est-à-dire le

carré du biais (théorique) plus la variance (théorique), où ce biais et cette variance se calculent sous la loi de probabilité donnée par le plan sur les facteurs non contrôlés en utilisation normale.

Exemple 13.2 *On veut expérimenter* 5 *facteurs : A, B, C, D et E à l'aide de* $32 = 2^5$ *unités, ce qui correspond à un plan complet. Malheureusement, on considère qu'il est impossible de réaliser plus de* 8 *unités de manière homogène, de sorte qu'il faut introduire un facteur bloc BL à* 4 *niveaux qui sera codé par deux pseudo-facteurs* B_1 *et* B_2*. On est donc amené à chercher un plan* 2^{7-2}*. Celui est donné par la table 4 ci-dessous, avec* $B_1 = A \cdot B \cdot C \cdot D$ *et* $B_2 = A \cdot B \cdot D \cdot E$*, ce qui amène à confondre l'interaction* $C \cdot E$ *avec* $B_1 \cdot B_2$ *qui est un effet bloc. On laisse à titre d'exercice, le soin au lecteur de vérifier que le plan suivant :* $B_1 = A \cdot B \cdot C$ *et* $B_2 = C \cdot D \cdot E$ *est meilleur.*

5 Exemples traités par logiciels informatiques

Construction de plans fractionnaires

Supposons que nous voulions construire un plan pour 5 facteurs : A, B, C, D et E, de résolution V avec 16 unités et la clef $E = A \cdot B \cdot C \cdot D$. La mise en pratique est la suivante :

Logiciel Splus :

La génération d'un tel plan fractionnaire avec le logiciel Splus peut être obtenue directement par la commande :

```
menuFacDesign(c(2,2,2,2,2),factor.names=c("a","b","c","d","e"),
                                          fraction=~A:B:C:D:E)
```

La clef est ici $A \cdot B \cdot C \cdot D \cdot E$, puisque $A \cdot B \cdot C \cdot D \cdot E = 1$. Le résultat est une matrice à 16 lignes, qui correspondent aux unités, et 5 colonnes qui correspondent aux facteurs. En rajoutant dans la commande l'option `randomize.rows=T`, on randomise les lignes (ce qui revient à faire une permutation des unités) et on obtient alors (ici les différents niveaux de a sont notés $a1$ et $a2$ et non -1 et 1) :

```
a1  b1  c1  d2  e1
a2  b2  c1  d2  e1
a2  b2  c2  d1  e1
a1  b2  c1  d1  e1
a2  b1  c1  d1  e1
:   :   :   :   :
```

Notons que cette commande `menuFacDesign` est extrêmement puissante, puisque l'on peut également choisir une fraction quelconque du plan, et des variables avec plus de deux niveaux.

Logiciel R :

La génération d'un tel plan fractionnaire avec le logiciel R peut être obtenue par la suite de commandes suivantes, après avoir fait appel au module `AlgDesign` :

Nombre unités	Plans fractionnaire pour p facteurs à deux niveaux Nombre de facteurs p								
	3	4	5	6	7	8	9	10	11
4	III 2^{3-1} C=A · B	/	/	/	/	/	/	/	/
8	Plan Complet	IV 2^{4-1} D=ABC	III 2^{5-2} D=A · B E=A · C	III 2^{6-3} D=A · B E=A · C F=B · C	III 2^{7-4} D=A· B E=A· C F=B· C G=A· B· C	/	/	/	/
16		Plan Complet	V 2^{5-1} E=A· B· C· D	IV 2^{6-2} E=A· B· C F=B· C· D	IV 2^{7-3} E=A· B· C F = B· C· D G = A· C· D	IV 2^{8-4} E=B· C· D F=A· C· D G=A· B· C H=A· B· D	III 2^{9-5} E=A· B· C F=B· C· D G=A· C· D H=A· B· D I=A· B· C· D	III 2^{10-6} E=A· B· C F=B· C· D G=A· C· D H=A· B· D I=A· B· C· D J=A· B	III 2^{11-7} E=A· B· C F=B· C· D G=A· C· D H=A· B· D I=A· B· C· D J=A· B K =A· C
32			Plan Complet	VI 2^{6-1} F=A· B C· D· E	IV 2^{7-2} F=A· B· C· D G=A· B· D· E	IV 2^{8-3} F=A· B· C G=A· B· D H=B· C· D· E	IV 2^{9-4} F=B· C· D· E G=A· C· D· E H=A· B· D· E I=A· B· C· E	IV 2^{10-5} F=A· B· C· D G=A· C· D· E H=A· B· D· E I=A· C· D· E J=B· C· D· E	IV 2^{11-6} F=A· B· C G=B· C· D H=C· D· E I=A· C· D J=A· D· E K =B· D· E
64				Plan Complet	VII 2^{7-1} G=A· B· C D· E· F	V 2^{8-2} G=A· B· C· D H=A· B· E· F	IV 2^{9-3} G=A· B· C· D H=A· C· E· F I=C· D· E· F	IV 2^{10-4} G=B· C· D· F H=A· C· D· E I=A· B· D· E J=A· B· C· E	IV 2^{11-5} G=C· D· E H=A· B· C· D I=A· B· F J=B· D· E· F K=A· D· E· F
128					Plan Complet	VIII 2^{8-1} H=A· B· C· D· E· F· G	VI 2^{9-2} H=A· C· D F· G I=B· C· E· F· G	V 2^{10-3} H=A· B· C· G I=B· C· D· E J=A· C· D· F	V 2^{11-4} H=A· B· C· G I=B· C· D· E J=A· C· D· F K=A· B· C· D· E· F· G

TABLE 13.1 – Tableau de plans fractionnaires de minimum d'aberration, avec, en ligne, le nombre d'unités 2^{p-q} et en colonne le nombre de facteurs. Une solution est donnée lorsqu'il existe un plan de résolution III au moins. Chaque cellule indique (de haut en bas) : la résolution, le nom du plan, les clefs.

```
library(AlgDesign)
plan1=gen.factorial(2,4,varNames=c("a","b","c","d"))
attach(plan1)
e=a*b*c*d
plan=data.frame(plan1,e)
obs=sample(1:16)
plan=data.frame(as.matrix(plan)[obs,1:5])
```

Notons que la commande `gen.factorial` permet la construction d'un plan complet. Les deux dernières commandes permettent la randomisation ***** des colonnes du plan d'expériences.

Logiciel SAS :

La procédure `proc plan` permet de construire des plans par exemple en utilisant la table 4 comme nous allons le décrire. La construction de plans pour facteurs à 3 niveaux, l'étude exhaustive des confusions est possible en utilisant la procédure `proc factex` du module spécialisé SAS/QC. Comme sa diffusion est plus confidentielle, nous nous contenterons de mentioner l'existence de cette dernière procédure.

```
proc plan;
factors a=2 b=2 c=2 d=2;
output out=sortie a nvals=(-1 1) b nvals=(-1 1) c nvals=(-1 1) d nvals=(-1 1);
run;quit;
```

qui donne le plan factoriel complet pour les 4 facteurs de base. On introduit le 5-ème et la clef par :

```
data planfrac; set sortie; e=a*b*c*d; proc print; run;
```

La plan se trouve dans le data "planfrac".
Voici un résultat possible (en R ou SAS) :

Obs	a	b	c	d	e
1	1	1	-1	-1	1
2	1	1	-1	1	-1
3	1	1	1	-1	-1
4	1	1	1	1	1
5	1	-1	-1	1	1
:	:	:	:	:	:

6 Exercices

Exercice 13.1

(*) En reprenant la définition 13.2, construisez les différents espaces :

$$V_\emptyset,\ V_{\{1\}},\ V_{\{2\}},\ V_{\{1,2\}},\ W_\emptyset,\ W_{\{1\}},\ W_{\{2\}},\ W_{\{1,2\}}$$

dans le cas de $p = 3$ facteurs.

Exercice 13.2

(*) Choisir un plan permettant d'expérimenter 7 facteurs $A, B, \cdots, G$, en estimant les effets principaux en présence d'éventuelles interactions. Donner les relations complètes de définitions et l'alias du facteur A.

Exercice 13.3

(**) On veut construire un plan pour 5 facteurs A, B, C, D et E, tel que l'on veut pouvoir estimer sans confusion les effets principaux et les interactions suivantes : $A*D$, $A*E$, $B*D$ et $B*E$. Proposez une solution de dimension minimale. Même question avec seulement $A*D$, $A*E$.

Exercice 13.4

(**) On cherche un plan pour A à 4 niveaux codés par deux pseudo-facteurs A_1 et A_2 et 5 facteurs B, C, D, E, F à deux niveaux. A partir de la table 4, proposez un quart de plan. Calculez avec soin sa résolution en prenant en compte l'existence de pseudo-facteurs. Peut-on faire mieux ?

Exercice 13.5

(*) Combien faut-il d'unités au moins pour expérimenter 12 facteurs en résolution IV et V respectivement ?

Exercice 13.6

(*) [Repliement de plans] Un expérimentateur réalise le plan 2^{5-2} défini par les relations $D = A \cdot B \cdot C$ et $E = B \cdot C$. Plus tard, il réalise un plan identique, sauf que tous les signes ont été inversés.

i. Quelle est la résolution du plan ainsi défini ?

ii. Aurait-on choisi ce plan si, dès le départ, on avait su disposer de 16 unités ?

Exercice 13.7

(**) Soit A un facteur à 4 niveaux et B, C, D, E, 4 facteurs à deux niveaux. Pour ces 5 facteurs on désire construire un plan à 32 unités en deux blocs de longueur 16. Les blocs sont numérotés par un facteur BL. On suppose qu'il n'y a pas d'interaction d'ordre 3 ou plus et pas d'interaction d'ordre 2 comprenant le facteur BL. Proposer un plan dans lequel tous les effets principaux et toutes les interactions (ne comprenant pas BL) sont estimables.

Exercice 13.8

(***) [Cas d'un facteur à 3 niveaux] Supposons que l'on veuille étudier un facteur A à 3 niveaux et 3 facteur B, C, D à deux niveaux et réaliser un demi-plan. On peut montrer théoriquement qu'il est très difficile de mélanger des facteurs ayant des nombres de niveaux dont les décompositions en nombres premiers est différentes. Construire un plan 2^{3-1} pour B, C, D et combiner avec les 3 niveaux de A. On obtient un plan à 24 unités. Ajouter un niveau à A, coder le par deux pseudo-facteurs $A1$ et $A2$ et proposez un plan 2^{5-1} pour $A1, A2, B, C, D$. Enlever le niveau ajouté : on obtient bien un plan à 24 unités. Informatiquement, créez les matrices d'information des 2 plans et comparez les pour des critères de trace et de déterminant. Voir chapitre 14 pour une description détaillée de ces critères.

Chapitre 14

Surfaces de réponses et plans isovariants

Ce chapitre étudie la planification d'expériences pour un certain nombre de variables explicatives quantitatives pouvant répondre à certaines contraintes et intervenant dans un modèle linéaire fixé a priori. Ceci amène à définir les plans isovariants, dont le précision est invariante par isométrie, puis à étudier les modèles polynômiaux. Le cas particulier des modèles quadratiques permet la construction explicite d'une solution : les plans composites centrés de Box et Wilson. Dans les cas plus complexes, différentes notions d'optimalité sont proposées, ainsi que leurs solutions numériques.

1 Cadre de l'étude

1.1 Un exemple

On considère une expérience de chimie dans laquelle on veut étudier le rendement R d'une réaction chimique, c'est-à-dire la proportion de réactif transformé. On suppose que cette variable peut être influencée par deux variables : le pH, PH et la température T de la solution. On suppose de plus que ces deux variables peuvent varier librement dans un certain domaine. On choisit plusieurs conditions expérimentales (T_i, PH_i) pour $i = 1, \cdots, n$ et on obtient ainsi $(R_i)_{1 \leq i \leq n}$. On pose a priori comme modèle un modèle de régression quadratique

$$R_i = \mu + \beta_1 \cdot T_i + \beta_2 \cdot PH_i + \beta_3 \cdot T_i \cdot PH_i + \beta_4 \cdot T_i^2 + \beta_5 \cdot PH_i^2 + \varepsilon_i \quad \text{pour} \quad i = 1, \ldots, n,$$

où $\mu, \beta_1, \ldots, \beta_5$ sont les paramètres réels inconnus et (ε_i) une erreur gaussienne centrée de variance σ^2. En considérant X la matrice dont les colonnes sont : $\mathbb{1}$, $(T_i)_i$, $(PH_i)_i$ $(T_i \cdot PH_i)_i$ $(T_i^2)_i$ et $(PH_i^2)_i$ (dans $\mathbb{R}^n$), on obtient le modèle linéaire :

$$R = X \cdot \theta + \varepsilon,$$

avec $R = (R_i)$, $\theta = (\mu, \beta_1, \beta_2, \beta_3)'$ et $\varepsilon = (\varepsilon_i)_i$. Après avoir obtenu les estimateurs $\widehat{\theta}$, on se demande quelle valeur prendrait R si faisait l'expérience en une nouvelle condition $(T, PH) = (t, ph)$, où (t, ph) est expérimentalement réalisable.

1.2 Cas général

Retournons à un cadre plus général. On considère une variable Y (dans l'exemple R) éventuellement influencée par m variables quantitatives réelles $Z^{(1)}, \cdots, Z^{(m)}$ (dans l'exemple T et PH). On suppose que l'on dispose de n réalisations du vecteur $Z = (Z^{(1)}, \cdots, Z^{(m)})$, notées $(Z_i^{(1)}, \cdots, Z_i^{(m)})_{1 \leq i \leq n}$ et que l'on a pu observer les réponses $(Y_i)_{1 \leq i \leq n}$. On suppose également que :

- $\mathbb{E}(Y)$ est une fonction polynômiale de degré q de $Z^{(1)}, \cdots, Z^{(m)}$;
- Toutes les observations Y sont non-corrélées et de même variance (même si les conditions expérimentales $(Z^{(1)}, \cdots, Z^{(m)})$ peuvent être répétées plusieurs fois).

En ordonnant arbitrairement les unités du plan, on obtient un modèle linéaire classique non-gaussien.

$$\mathbb{E}(Y) = X \cdot \theta \quad \text{et} \quad \mathrm{Var}(Y) = \sigma^2 \cdot Id. \tag{14.1}$$

On note k la dimension de θ et donc, également le nombre de colonnes de la matrice $X = (X_{ij})_{1 \leq i \leq n, 1 \leq j \leq k}$, matrice dont les coordonnées vérifient pour $1 \leq i \leq n$ et pour $1 \leq j \leq k$,

$$X_{ij} = \left(Z_i^{(1)}\right)^{\gamma_{j1}} \times \cdots \times \left(Z_i^{(m)}\right)^{\gamma_{jm}}, \quad \text{avec} \quad \gamma_{jl} \in \mathbb{R} \quad \text{et} \quad \sum_{l=1}^{k} \gamma_{jl} \leq q. \tag{14.2}$$

Dans l'exemple plus haut, $X_{i1} = T_i^2$, $X_{i2} = PH_i$ et $X_{i3} = T_i^2 \cdot PH_i$. Certaines colonnes de la matrice X sont liées fonctionnellement : certaines sont des produits ou des puissances d'autres colonnes. Nous supposerons cependant qu'elles ne sont pas liées linéairement de sorte que le modèle (14.1) est régulier.

Dans toute la suite, on note $z := (z_1, \cdots, z_m) \in \mathbb{R}^m$ une valeur possible du vecteur $(Z^{(1)}, \cdots, Z^{(m)})$, qui peut aussi bien être une des réalisations déjà

effectuée $(Z_i^{(1)}, \cdots, Z_i^{(m)})$, qu'une valeur différente. On notera également $\widetilde{X}(z)$ le vecteur ligne de longueur k tel que :

$$\widetilde{X}(z) = \widetilde{X}(z_1, \cdots, z_m) = \left(\prod_{l=1}^{m} (z_l)^{\gamma_{1l}}, \cdots, \prod_{l=1}^{m} (z_l)^{\gamma_{kl}} \right), \tag{14.3}$$

en utilisant les γ_{jl} définis précédemment. $\widetilde{X}(z)$ correspondrait à la ligne de X ou se situeraient les composantes de z, si l'observation z avait été faite.

Définition 14.1 *Soit $\widehat{\beta}$ l'estimateur par moindres carrés de β, soit $\widehat{\beta} = (X' \cdot X)^{-1} \cdot X' \cdot Y$. On dit que le plan est isovariant (rotatable en anglais) si pour tout $z \in \mathbb{R}^m$, la fonction $Var\left(\widetilde{X}(z) \cdot \widehat{\beta}\right)$ ne dépend de z qu'a travers $\|z\|$.*

2 Conditions d'isovariance

Pour examiner les conséquences de la définition, commençons par énoncer deux lemmes d'algèbre linéaire utiles par la suite :

Lemme 14.1 *Si P est une transformation orthogonale de $\mathbb{R}^m$ de matrice P, il existe un endomorphisme (ou une matrice) unique Q_P sur $\mathbb{R}^k$ tel que pour tout $z \in \mathbb{R}^m$,*

$$Q_P \cdot \widetilde{X}'(z) = \widetilde{X}'(P \cdot z). \tag{14.4}$$

De plus, comme $P' = P^{-1}$, on a $Q_{P'} = (Q_P)^{-1}$.

Lemme 14.2 *Soit A une matrice carrée symétrique définie positive de taille p. Alors :*

$$\left[Tr(A) = Tr(A^{-1}) = p \right] \iff [A = Id].$$

Nous laissons la démonstration de ces deux lemmes à titre d'exercice.

Soit M la matrice dite des moments, qui est la matrice d'information (qui sera, elle, notée N) divisée par le nombre de données :

$$M := \frac{1}{n} \cdot X' \cdot X = \frac{1}{n} \cdot N.$$

On peut maintenant caractériser un plan isovariant à partir de cette matrice M des moments du modèle :

Proposition 14.1 *Le plan est isovariant si et seulement si M est isovariante c'est-à-dire que pour toute transformation orthogonale P, on a $M = Q_P \cdot M \cdot Q'_P$.*

Démonstration : **1.** Montrons d'abord le sens $\Longleftarrow$. Supposons donc que M est une matrice isovariante. On veut montrer que le plan est isovariant. Pour cela, pour P transformation orthogonale, on calcule :

$$\begin{aligned}
\mathrm{Var}\left(\widetilde{X}\,(P\cdot z)\cdot\widehat{\beta}\right) &= \widetilde{X}\,(P\cdot z)\cdot\mathrm{Var}\,(\widehat{\beta})\cdot\widetilde{X}'\,(P\cdot z)\\
&= \widetilde{X}(z)\cdot Q'_P\cdot\mathrm{Var}\,(\widehat{\beta})\cdot Q_P\cdot\widetilde{X}(z)' \qquad \text{(d'après le lemme 14.1)}\\
&= \sigma^2\cdot\widetilde{X}(z)\cdot Q'_P\cdot(X'\cdot X)^{-1}\cdot Q_P\cdot\widetilde{X}(z)' \qquad \text{(calcul de la variance de } \widehat{\beta})\\
&= \sigma^2\cdot\widetilde{X}(z)\cdot(X'\cdot X)^{-1}\cdot\widetilde{X}(z)' \qquad \text{(car la matrice } M \text{ est isovariante)}\\
&= \mathrm{Var}\left(\widetilde{X}(z)\cdot\widehat{\beta}\right).
\end{aligned}$$

Comme ce résultat est vrai pour toute transformation orthogonale P, on en déduit que $\mathrm{Var}\left(\widetilde{X}(z)\cdot\widehat{\beta}\right)$ ne dépend donc que de $\|z\|$: le plan est bien isovariant.

2. Montrons maintenant le sens $\Longrightarrow$. Nous supposons donc que le plan est isovariant. Remarquons d'abord que

$$\begin{aligned}
\mathrm{Tr}(\mathrm{Var}\,(X\cdot\widehat{\beta})) &= \sigma^2\cdot\mathrm{Tr}(X\cdot(X'\cdot X)^{-1}\cdot X')\\
&= \sigma^2\cdot\mathrm{Tr}((X'\cdot X)^{-1}\cdot X'\cdot X)\\
&= \sigma^2\cdot k
\end{aligned}$$

(en utilisant le fait que $\mathrm{Tr}(A\cdot B)=\mathrm{Tr}(B\cdot A)$). Soit maintenant le plan transformé par P, une transformation orthogonale quelconque. L'équation (14.4) peut être utilisée pour montrer que la matrice de ce plan vaut $X\cdot Q'_P$. Le plan étant isovariant,

$$\mathrm{Var}\left(X\cdot Q'_P\cdot\widehat{\beta}\right)=\mathrm{Var}\left(X\cdot\widehat{\beta}\right).$$

Donc en utilisant ce qui précède, on a :

$$\begin{aligned}
\sigma^2\cdot k &= \mathrm{Tr}\big(\mathrm{Var}\left(X\cdot\widehat{\beta}\right)\big)\\
&= \mathrm{Tr}\big(\mathrm{Var}\left(X\cdot Q'_P\cdot\widehat{\beta}\right)\big)\\
&= \sigma^2\cdot\mathrm{Tr}\big(X\cdot Q'_P\cdot(X'\cdot X)^{-1}\cdot Q_P\cdot X'\big)\\
\text{soit}\quad k &= \mathrm{Tr}\big(Q'_P\cdot(X'\cdot X)^{-1}\cdot Q_P\cdot X'\cdot X\big)\\
&= \mathrm{Tr}\big(Q'_P\cdot M^{-1}\cdot Q_P\cdot M\big)
\end{aligned}$$

Posons : $A:=\big(M^{1/2}\cdot Q'_P\cdot M^{-1}\cdot Q_P\cdot M^{1/2}\big)$. On vient de voir que

$$\mathrm{Tr}(A)=k.$$

Par ailleurs, on a :

$$A^{-1}=\big(M^{-1/2}\cdot Q_P^{-1}\cdot M\cdot(Q'_P)^{-1}\cdot M^{-1/2}\big).$$

Donc :

$$\begin{aligned} \mathrm{Tr}(A^{-1}) &= \mathrm{Tr}\big((Q'_P)^{-1} \cdot M^{-1} \cdot Q_P^{-1} \cdot M^{-1}\big) \\ &= k \quad \text{car on a vu que } \; Q_P^{-1} = Q_{P'}. \end{aligned}$$

En appliquant le lemme 14.2, on montre que A et A^{-1} valent l'identité, ce qui implique que $M = Q_P^{-1} \cdot M \cdot (Q'_P)^{-1}$, soit $M = Q_{P'} \cdot M \cdot (Q'_{P'})$ pour toute transformation orthogonale P : la matrice M est bien isovariante. ■

Nous en venons maintenant au résultat principal de cette section. En raison du modèle polynomial et de l'écriture des X_{ij} donnée en (14.2), les éléments de M s'écrivent tous sous la forme :

$$[\delta_1, \cdots, \delta_m] := \frac{1}{n} \sum_{i=1}^{n} (Z_i^{(1)})^{\delta_1} \times \cdots \times (Z_i^{(m)})^{\delta_m},$$

avec la condition $\delta_1 + \cdots + \delta_m \leq 2q$.

Théorème 14.1 (Box et Hunter 1957) *Avec les notations précédente, la matrice des moments M est isovariante si et seulement si une composante quelconque $[\delta_1, \cdots, \delta_m]$ de M de degré $\delta = \delta_1 + \cdots + \delta_m \leq 2q$,*

1. *est nulle si l'un des δ_j est impair ;*

2. *est égale à $\dfrac{\mu_\delta}{2^{\delta/2}} \displaystyle\prod_{j=1}^{m} \frac{(\delta_j)!}{(\delta_j/2)!}$ sinon, avec μ_δ des constantes réelles ne dépendant que de δ et avec la convention $0! = 1$.*

Avant d'entamer la démonstration de ce théorème, illustrons son contenu avec l'exemple incontournable des modèles quadratiques, qui correspondent à $q = 2$. Dans ce cas, on doit considérer les moments jusqu'à l'ordre 4. Compte tenu du point 1., les moments (composantes de la matrice M) non nuls ne peuvent être que ceux qui correspondent à :

- un δ_j égal à 2, les autres $\delta_{j'}$ étant tous nuls (donc $\delta = 2$). Tous les termes de la matrice M de degré 2 en les $Z_i^{(j)}$ sont tous égaux et d'après le théorème, on note $\dfrac{\mu_2}{2^1} \cdot \dfrac{2!}{1!} = \mu_2$ leur valeur commune ;
- deux δ_j égaux à 2, les autres étant tous nuls (donc $\delta = 4$). D'après le théorème, on note $\dfrac{\mu_4}{2^2} \cdot \dfrac{2! \cdot 2!}{1! \cdot 1!} = \mu_4$ la valeur commune de tous ses moments ;

- un δ_j égal à 4, les autres étant tous nuls. Ces moments là sont tous égaux et leur valeur est maintenant $\frac{\mu_4}{2^2} \cdot \frac{4!}{2!} = 3 \cdot \mu_4$.

Nous reviendrons sur cet exemple ultérieurement. Voici maintenant la démonstration du théorème.

<u>*Démonstration*</u> : Pour $t = (t_1, \cdots, t_m) \in \mathbb{R}^m$, on pose

$$F(t) := \frac{1}{n} \sum_{i=1}^{n} \left(1 + Z_i^{(1)} t_1 + \cdots + Z_i^{(m)} t_m\right)^{2q}.$$

En utilisant la formule du binôme généralisée, on montre que pour $\delta \leq 2q$, le coefficient du terme en $t_1^{\delta_1} \times \cdots \times t_m^{\delta_m}$, vaut :

$$\frac{(2q)!}{(2q-\delta)! \prod_{j=1}^{m} (\delta_j)!} [\delta_1, \cdots, \delta_m].$$

Supposons la matrice des moments M isovariante, les éléments ci-dessus sont alors invariants par isométrie. Cela implique que pour toute transformation orthogonale P :

$$F(P(t)) = F(t).$$

La fonction F est donc une fonction de la norme (euclidienne classique) de t, et comme F est un polynôme, F est un polynôme en $\|t\|^2$ [1](ce qui montre que l'on ne peut avoir de δ_j impairs dans son expression), et on peut écrire

$$F(t) = \sum_{\delta=0}^{q} a_\delta (\|t\|^2)^\delta$$

où les a_δ sont certaines constantes.

$$F(t) = \sum_{\delta=0}^{q} a_\delta \left[\sum_{\delta_1 + \cdots + \delta_m = \delta} \frac{\delta!}{\prod_{j=1}^{m} (\delta_j)!} \times t_1^{2\delta_1} \times \cdots \times t_m^{2\delta_m} \right],$$

en utilisant à nouveau la formule du binôme généralisée. En identifiant ce développement de F avec celui présenté plus haut, on obtient le théorème en posant $\mu_{2\delta} = a_\delta \cdot \frac{2^\delta \cdot (2q-2\delta)! \cdot \delta!}{(2q)!}$. ■.

1. En toute rigueur, ce dernier énoncé n'est pas évident, mais on peut en trouver une démonstration détaillée dans le livre de Spivak [56]

3 Plans composites centrés de Box et Wilson

Nous allons concentrer notre attention sur les modèles quadratiques ($q = 2$) pour m facteurs. Un plan composite est constitué de trois parties :

- une partie factorielle composée du plan factoriel complet $\{-f, +f\}^m$ où f est un certaine constante. Comme le plan est défini à une homothétie on supposera, pour normaliser le problème, que $f = 1$ (on peut montrer qu'en fait une fraction régulière de résolution V suffit) ;

- une partie axiale composée de $2m$ points à distance α de l'origine et situés sur les axes (α reste à déterminer) ;

- une partie centrale composée de n_0 points au centre 0 (n_0 reste à déterminer).

La distribution du plan est invariante :

- par toute symétrie par rapport à un hyperplan d'équation $z_j = 0$;

- par toute permutation des coordonnées.

Deux conséquences immédiates de ces propriétés sont le fait que $[\delta_1, \cdots, \delta_m]$ est nul dès qu'un des δ_j est impair et que $[\delta_1, \cdots, \delta_m] = [\delta_{\sigma(1)}, \cdots, \delta_{\sigma(m)}]$ pour toute permutation (bijection) σ de $\{1, \cdots, m\}$ dans $\{1, \cdots, m\}$.

Soit F le nombre de points de la partie factorielle du plan, on déduit de ce qui précède que la matrice d'information $X' \cdot X$ vérifie (en notant dans l'ordre : la constante, les

effets linéaires, les carrés et enfin les produits) :

$$X' \cdot X = \begin{bmatrix} n & 0 & \dots & \dots & 0 & b & \dots & \dots & b & 0 & \dots & \dots & 0 \\ 0 & b & 0 & \dots & 0 & 0 & \dots & \dots & 0 & 0 & \dots & \dots & 0 \\ \vdots & 0 & \ddots & & \vdots & \vdots & \ddots & & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 & \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & b & 0 & \dots & \dots & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & 0 & c & d & \dots & d & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots & d & \ddots & & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots & \vdots & & \ddots & d & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \dots & 0 & d & \dots & d & c & 0 & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & 0 & 0 & \dots & \dots & 0 & a & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots & \vdots & \ddots & & \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots & \vdots & & \ddots & \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \dots & \dots & 0 & 0 & \dots & \dots & 0 & 0 & \dots & \dots & a \end{bmatrix} \tag{14.5}$$

avec

$$\begin{aligned} b: &= \sum_{i=1}^{n} (Z_i^{(1)})^2 = F + 2\alpha^2; \\ d: &= a := \sum_{i=1}^{n} \left(Z_i^{(1)} Z_i^{(2)}\right)^2 = F; \\ c: &= \sum_{i=1}^{n} (Z_i^{(1)})^4 = F + 2\alpha^4. \end{aligned}$$

$\Longrightarrow$ **Le plan est isovariant si** $\alpha = F^{1/4}$**.**

On peut remarquer que pour $m = 2$ et $m = 4$, les points de la partie axiale et ceux de la partie factorielle sont situés sur la même sphère. Pour les autres valeurs de m, les points n'ont seulement qu'approximativement la même norme.

La Figure 3 permet de visualiser les plans composites centrés de Box et Wilson dans $\mathbb{R}^2$ et dans $\mathbb{R}^3$. Plus généralement, le Tableau 2 fournit un récapitulatif des possibilités de plans composites centrés suivant le nombre de variables considérées :

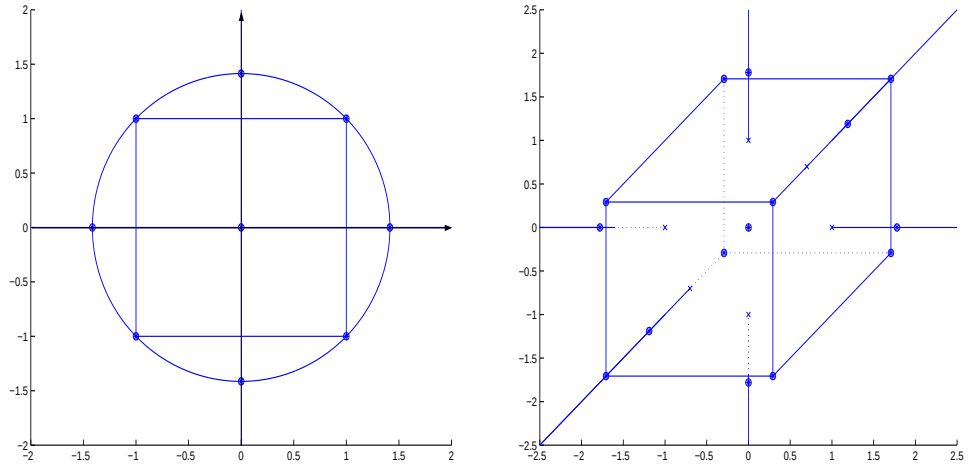


FIGURE 14.1 – Choix des points du plan d'expérience composite centré de Box et Wilson dans $\mathbb{R}^2$ (à gauche) et dans $\mathbb{R}^3$ (à droite).

Nombre de variables m	2	3	4	5	5	6	6
Taille p. fact. F	4	8	16	32	16	64	32
Taille p. axiale	4	6	8	10	10	12	12
Point centraux							
Isovariance	≥ 1	≥ 1	≥ 1	≥ 1	≥ 1	≥ 1	≥ 1
α	1.41	1.68	2	2.38	2	2.83	2.38

Tableau 2 : Récapitulatif des paramètres et des tailles des plans composites centrés.

Il existe d'autre méthodes de construction de plans isovariants. Certaines de ces méthodes sont basées sur des polygones réguliers. On trouvera une table de tels plans à l'adresse suivante : `http://www.dunod.com`.

4 Plans optimaux

Les plans isovariants ci-dessus ont, du fait de la propriété même d'isovariance, des propriétés d'optimalité que nous ne détaillerons pas. Cependant il existe de nombreux cas dans lesquels le domaine d'étude n'est pas "sphérique" et où les plans "composites centrés" ne peuvent pas être utilisés.

Exemple : On désire minimiser la production de polluants d'un moteur diesel mo-

derne à injection multiple. Le calculateur embarqué peut, pour un régime moteur donné et une puissance déterminée par l'utilisateur, faire varier les paramètres suivants :

- l'instant de la première injection, notée par les professionnels en bon français : SOIPREV (start of injection, previous) ;
- la masse de combustible injectée MFPREV (mass of fuel, previous) ;
- l'instant de la seconde injection SOIMAIN (start of injection, main) ;
- le taux de recyclage des gaz ;
- la pression d'injection.

Chaque variable varie a priori dans un certain domaine de sorte que le domaine de l'étude est "rectangulaire", sauf qu'il doit y avoir une distance minimale entre les deux injections :

$$\text{SOIMAIN-SOIPREV} > 15^0, \tag{14.6}$$

où la mesure est donnée en degrés d'angles correspondant à la rotation du vilebrequin. Le domaine de variation n'est alors ni rectangulaire ni sphérique et la solution du problème de l'isovariance n'est pas connue.

Il nous faut définir ce qu'est un "bon plan d'expériences" dans un cadre très général. Une façon de faire est de demander que la matrice de variance-covariance des estimateurs $\widehat{\theta}$ soit petite. Cependant il n'existe pas de "bonne" relation d'ordre entre deux matrices de variance-covariance. La plus naturelle est donnée par l'ordre de Löwner : la matrice M est dite supérieure à la matrice N si et seulement si $M - N$ est semi-définie positive. Malheureusement avec cet ordre (qui est dit partiel) peu de matrices sont comparables (condition très restrictive) et l'optimalité pour l'ordre de Löwner ne peut être obtenue que dans des cas très particuliers. On va donc rechercher des critères plus faibles mais qui permettent d'obtenir des plans optimaux dans tous les cas. Les critères que l'on peut considérer sont, par ordre d'importance :

- la D-optimalité : il s'agit de minimiser (en négligeant le facteur σ^2 qui ne joue aucun rôle)

$$\min\left\{ \det\left((X' \cdot X)^{-1}\right)\right\} = \frac{1}{\max\left\{ \det(X' \cdot X)\right\}}.$$

  Cela revient donc à maximiser le déterminant la matrice d'information. Cela revient à minimiser le volume de l'ellipsoïde de confiance pour $\widehat{\theta}$. Pour plus de

détail on pourra consulter par exemple [23]

– la E-optimalité : il s'agit de minimiser la plus grande valeur propre de la matrice de variance-covariance :
$$\min\Big\{\rho(X'\cdot X^{-1})\Big\}.$$
(on désigne souvent par ρ le rayon spectral de la matrice, c'est-à-dire sa plus grande valeur propre).

– la A-optimalité : cette fois-ci, on minimise
$$\min\Big\{\mathrm{Tr}\big((X'\cdot X)^{-1}\big)\Big\}.$$

Pour des problèmes de régression sur la droite réelle ou pour des plans qui sont des produits de plans pour chacune des variables, les solutions optimales sont connues (voir [23]). Pour les plans d'expériences de type blocs complets, plans en blocs incomplets équilibrés, on peut également montrer que ces choix sont optimaux pour les trois critères ci-dessus (voir par exemple [36], [37] ou [25]). Pour les autre cas, il n'y a pas de solution explicite mais seulement des algorithmes itératifs d'optimisation. En particulier les procédures `optex` du logiciel SAS, `optBlock` et `optFederov` du logiciel R donnent la possibilité de trouver la configuration optimale à partir d'un plan initial.

5 Exemples traités par logiciels informatiques

On reprend l'exemple des moteurs diesels en oubliant la contrainte (14.6), mais on ne donnera pas de commandes en Splus car l'optimisation de surface de réponse n'existe pas pour ce logiciel (dans la version Splus 2000 dont nous disposons).

Logiciel SAS :

Voici la suite de commande SAS permettant de rechercher un plan optimal :

```
proc plan;
factors a=4 b=4 c=4 d=4 e=4;
output out=table33 a nvals=(-1 -0.33 0.33 1) b nvals=(-1 -0.33 0.33 1)
c nvals=(-1 -0.33 0.33 1) d nvals=(-1 -0.33 0.33 1) e nvals=(-1 -0.33 0.33 1);
run;quit;
```

Cet appel de la procédure `proc plan` crée un plan initial pour 5 variables $a-e$ dont les

valeurs ont toutes été normalisées à 4 valeurs uniformément réparties sur l'intervalle $[-1, 1]$. Ce plan initial comprend donc $4^5 = 1024$ points : c'est beaucoup trop. On va donc rechercher un plan D-optimal pour un modèle cubique (rentré "à la main" de façon un peu laborieuse...) parmi ces 1024 possibilités :

```
proc optex data=table33;
model a b c d e a*a a*b a*c a*d a*e b*b b*c b*d b*e c*c c*d c*e d*d
d*e e*e a*a*a a*a*b a*a*c a*a*d a*a*e a*b*b a*b*c a*b*d a*b*e a*c*c
a*c*d a*c*e a*d*d a*d*e a*e*e b*b*b b*b*c b*b*d b*b*e b*c*c b*c*d
b*c*e b*d*d b*d*e b*e*e c*c*c c*c*d c*c*e c*d*d c*d*e c*e*e d*d*d
d*d*e d*e*e e*e*e ;
generate iter=15 n=70;
output out=cubique1; run;quit;
```

La recherche se fait parmi les unités du plan précédent. Le programme va chercher un plan de taille 70 (on a précisé $n = 70$ et si l'on ne spécifie rien, la taille est par défaut fixée à 10 plus la dimension du modèle) et donner les 15 meilleures solutions (`iter=15`). Les commandes ci-dessous permettent maintenant d'imprimer le plan.

```
proc print data=cubique1; run; quit;
```

Voici le résultat :

Design Standard Number	D-efficiency	A-efficiency	G-efficiency	Average Prediction Error
1	18.2758	4.1033	58.7749	1.1321
2	18.1228	3.5925	56.5411	1.1613
3	18.0057	3.7093	53.2260	1.1718
4	17.9982	4.1690	52.8823	1.1233
5	17.9885	3.8112	56.3177	1.1356
:	:	:	:	:

Les 15 meilleures solutions sont rangées selon le critère de D-optimalité. Le critère E est noté G par SAS. Notez le critère de l'erreur moyenne de prévision qui est parfois

très pertinent.
Voici enfin un aperçu des données (il y en a 70 en fait) correspondant à la "meilleure" solution (première pour la D-optimalité) :

OBS	A	B	C	D	E
1	-1.00	-1.00	-1.00	-1.00	0.33
2	-1.00	-1.00	-1.00	-0.33	-1.00
3	-1.00	-1.00	-1.00	1.00	-0.33
4	-1.00	-1.00	-1.00	1.00	1.00
5	-1.00	-1.00	-0.33	-1.00	1.00
6	-1.00	-1.00	1.00	-1.00	-1.00
:	:	:	:	:	:

Logiciel R :

Voici le même type de recherche de plan d'expériences "optimal" (utilisant l'algorithme de Federov) avec le logiciel R :

```
library(AlgDesign)
levels=c(-1,-0.33,0.33,1)
plan1=expand.grid(list(A=levels,B=levels,C=levels,D=levels,E=levels))
desL=optFederov(~cubic(.),plan1,nTrials=70,eval=TRUE)
```

Voici un extrait du plan sélectionné :

```
$D [1] 0.3139053

$A [1] 17.53834

$I [1] 33.85085

$Ge [1] 0.672

$Dea [1] 0.614

$design
         A     B     C     D     E
1    -1.00 -1.00 -1.00 -1.00 -1.00 4     1.00 -1.00 -1.00 -1.00
```

```
-1.00 13   -1.00  1.00 -1.00 -1.00 -1.00 16    1.00  1.00 -1.00
-1.00 -1.00 18   -0.33 -1.00 -0.33 -1.00 -1.00 49   -1.00 -1.00
1.00 -1.00 -1.00 :       :     :     :     : :       :     :
:     :     : 1013 -1.00 -0.33  1.00  1.00  1.00 1021 -1.00  1.00
1.00  1.00  1.00 1024  1.00  1.00  1.00  1.00  1.00
```

Ici l'optimisation est effectuée suivant le critère de D-optimalité (dont la définition diffère de la définition donnée précédemment par le fait que l'on minimise $\{\det\left((X' \cdot X)^{-1}\right)^{1/k}$, où k est le nombre de variables), mais l'on peut demander que cette optimisation soit faite également suivant un des autres critère (la A-optimalité diffère également de la définition donnée précédemment par un facteur multiplicatif en $1/k$ près).

6 Exercices

Exercice 14.1

(**) Soit le modèle de régression linéaire simple :

$$Y_i = \alpha + \beta \cdot Z_i + \varepsilon_i,$$

dans lequel les valeurs de la variable Z appartiennent à l'intervalle $[a, b]$, $a < b$. Montrer qu'un plan optimal comprenant $2n$ points consiste en n observations en a et n observations en b.

Exercice 14.2

(**) Montrer qu'effectivement un fraction de résolution V sur la partie factorielle suffit dans un plan composite centré.

Exercice 14.3

(**) Montrer que la dimension k du modèle polynomial de degré q en m variables vaut C^m_{m+q}.

Chapitre 15

Etudes de cas traités par logiciels informatiques

Pour illustrer une bonne part des chapitres précédents, compléter les courts exemples qui les illustrent et enfin montrer l'intérêt réel du modèle linéaire, on reprend ici trois études de cas qui ont été effectivement menées par les auteurs.

1 Etude de cas "Argus" : Analyse de la variance et de la covariance

1.1 Présentation

Le but de la présente étude de cas est d'essayer d'établir une modélisation du prix de certaines voitures neuves ou d'occasion en fonction de leurs âges et du modèle choisi. La source de nos données est la cote donnée par le journal "l'Argus de l'automobile" de certains modèles d'automobiles à la date du 1er juillet 2002. Les modèles retenus (variable `modele`) étaient ceux qui intéressaient l'un des auteurs à cette date, son choix entièrement subjectif a porté sur :

`bravae`	:	Fiat Brava 100sx (produite jusqu'en 2001)
`bravad`	:	Fiat Brava TD75 (produite jusqu'en 2001)
`mondeoe`	:	Ford Mondeo 1.8 GLX 5P
`mondeod`	:	Ford Mondeo Td GlX 5p
`a140`	:	Mercedes A140
`omega`	:	Opel Omega Elegance 2.2 (2.0 avant 2000)
`avensis`	:	Toyota Avensis 1.6 sol 5p
`megane`	:	Renault Mégane 1.4 RTE
`octavia`	:	Skoda Octavia 1.8 SLX

(les appellations commerciales sont celles de 2002, elles ont pu changer en fonction de l'évolution des modèles).

Les informations recueillies par ailleurs sont :

- l'âge en années de la voiture : ce sera la variable `age`, qui prendra les valeurs 1, 1.5, 2, 2.5, 3.5 et 4.5 (en années). Les demi-années sont dues à des obscures raisons de changement de référence dans l'année du modèle.

- le prix (variable `prix`) et son logarithme (variable `log`) qui est soit le prix neuf si `age`$= 0$, soit la "cote Argus" sinon.

1.2 Premières modélisations

Une première idée pour obtenir une modélisation serait de faire une analyse de la variance en expliquant le prix par un modèle additif à l'aide des variables `modele` et `age`, cette dernière variable étant déclarée comme qualitative. Comme il n'y a pas de répétition, $n_{ij} = 1$ et il n'est donc bien-sûr pas possible de rajouter une interaction.

Ce modèle est trivial et il ne va conclure qu'à la banalité suivante : le prix dépend de l'âge et du modèle.

Cette analyse peut être légèrement améliorée en travaillant sur le logarithme du prix, ce qui revient à poser un modèle multiplicatif sur le prix lui même. En effet, poser

$$log_{ij} \simeq \mu + age_i + modele_j$$

revient à supposer

$$prix_{ij} \simeq \tilde{\mu} \cdot \tilde{age}_i \cdot \tilde{modele}_j,$$

ce qui est plus proche de la réalité que le modèle de départ (l'intuition préfère une telle modélisation). En utilisant le logiciel SAS, on déclare ce modèle par :

```
proc glm data=sasuser.argus;
class modele age;
model log=age modele;
run; quit;
```

Ce modèle nous servira en fait uniquement de référence en terme de qualité d'ajustement. Les indicateurs statistiques classiques obtenus sont :

$$R^2 \simeq 0.989 \ ; \ \widehat{\sigma} \simeq 0.045 \ ; \ AIC \simeq -188.9 \ ; \ AICC \simeq -170.6 \ ; \ BIC \simeq -155.2$$

pour une moyenne de `log` de 9.19. Si on interprète le $\widehat{\sigma}$ comme une erreur moyenne de la modélisation sur le logarithme, cela ce ramène à une erreur de 4.6% sur le prix, c'est-à-dire encore en moyenne : 450 Euros. Notez que pour obtenir les critères AIC $AICC$ et BIC faut faire appel à `proc mixed` plutôt qu'à `proc glm`, ou faire les calculs à la main...

Attention ! il faut absolument demander à `proc mixed` de calculer la vraisemblance et non pas la vraisemblance restreinte ce qui ce fait par `proc mixed data =.. method=ml;`

Remarque : L'utilisation des critères de sélection de modèles AIC et BIC que nous avons détaillée au chapitre 9 dans le cadre de la régression se généralise naturellement à celui de l'analyse de la variance car un modèle d'analyse de la variance n'est après tout qu'un modèle de regression sur les variables indicatrices générées par les facteurs.

1.3 Un modèle plus complet

Un modèle plus interprétable pourrait être celui constitué par des modèles d'analyse de la covariance : modéliser linéairement le logarithme du prix revient à supposer une décroissance exponentielle de ce même prix en fonction de l'âge. Il serait alors possible de poser le modèle complet avec "interaction" (hétérogénéité des pentes). Mais ce modèle ne serait pas réaliste car il est bien connu que les modèles neufs sont sur-cotés par rapport à ceux d'occasion.

Pour prendre ce phénomène en compte, on crée la variable `neuf` qui vaut 1 si `age`$= 0$. Le programme SAS associé est alors le suivant :

```
proc glm data=sasuser.argus;
class modele neuf;
```

```
model log=age modele age*modele neuf;
lsmeans neuf output out=a r=res p=pred;
proc gplot data=a;
plot res*pred=modele;
run; quit;
```

On obtient la sortie suivant (extrait) :

```
Dependent Variable: log
 Source                  DF         Squares     Mean Square    F Value    Pr > F
  Model                  18      9.21446384      0.51191466     268.10    <.0001
  Error                  42      0.08019645      0.00190944
  Corrected Total        60      9.29466029

                  R-Square     Coeff Var      Root MSE      log Mean
                  0.991372      0.475149      0.043697      9.196516

 Source               DF     Type III SS     Mean Square    F Value    Pr > F
  age                  1      3.76957692      3.76957692    1974.18    <.0001
  modele               8      0.60059953      0.07507494      39.32    <.0001
  age*modele           8      0.03834856      0.00479357       2.51    0.0251
  neuf                 1      0.06785510      0.06785510      35.54    <.0001

                                   neuf      log LSMEAN
                                   0         9.17422022
                                   1         9.31000248
```

En lançant le même programme exactement sous la procédure `proc mixed` en demandant `method =ML`, on obtient $AIC = -191.6$ $AICC = -170.6$ et $BIC = -149.4$. Le coefficient R^2 est donc légèrement meilleur que le précédent, tout comme l'écart-type $\widehat{\sigma}$ est légèrement plus faible (donc meilleur). L'hétérogénéité des pentes est faiblement significative. Tous les autres effets sont clairement significatifs. Les indicateurs de choix de modèle AIC $AICC$ et BIC sont plutôt ambigus en ce qui concerne la comparaison avec l'analyse de la variance. Les valeurs `lsmeans` montrent que si le prix d'une voiture neuve est 100 le prix d'un voiture d'occasion d'age 0 est : $\simeq 100 \times \exp(9.31 - 9.17) \simeq 87.3$.

Il faut donc négocier une remise de 13% chez votre marchand de voiture neuve pour que l'offre soit comparable à celle de l'occasion.

Par ailleurs, une autre analyse, non reproduite ici, a montré qu'un terme quadratique `age`^{2} était non significatif.

Maintenant il faut un peu d'astuce pour obtenir des estimateurs des paramètres dans ce modèle qui mélange variables qualitatives et quantitatives.

```
proc glm data=sasuser.argus;
class modele neuf hg diesel;
model log=modele neuf age*modele/noint solution;
run;quit;
```

Voici le résultat :

Parameter		Estimate		Standard Error	t Value	Pr > |t|
modele	a140	9.83026485	B	0.03238713	303.52	<.0001
modele	avensis	9.89539943	B	0.03238713	305.53	<.0001
modele	bravad	9.58824096	B	0.04733974	202.54	<.0001
modele	bravae	9.60470917	B	0.04733974	202.89	<.0001
modele	megane	9.58235168	B	0.03238713	295.87	<.0001
modele	mondeod	9.91159640	B	0.03238713	306.04	<.0001
modele	mondeoe	9.81312974	B	0.03238713	302.99	<.0001
modele	octavia	9.75980162	B	0.03238713	301.35	<.0001
modele	omega	10.22595109	B	0.03238713	315.74	<.0001
neuf	0	-0.13578225	B	0.02277743	-5.96	<.0001
neuf	1	0.00000000	B	.	.	.
age*modele	a140	-0.20104571		0.01225565	-16.40	<.0001
age*modele	avensis	-0.20698334		0.01225565	-16.89	<.0001
age*modele	bravad	-0.20984650		0.01498799	-14.00	<.0001
age*modele	bravae	-0.23120715		0.01498799	-15.43	<.0001
age*modele	megane	-0.23280385		0.01225565	-19.00	<.0001
age*modele	mondeod	-0.21490115		0.01225565	-17.53	<.0001
age*modele	mondeoe	-0.22124995		0.01225565	-18.05	<.0001
age*modele	octavia	-0.21588987		0.01225565	-17.62	<.0001
age*modele	omega	-0.26390048		0.01225565	-21.53	<.0001

Du fait que la correction due à la variable `age` est nulle pour les modèles neufs, les prix moyen (en logarithme) comprennent la sur-cote des modèles neufs. Il faut leur enlever 0.1357 pour obtenir le prix en occasion... Plus importantes sont les pentes qui sont estimables au sens des modèles non réguliers, puisqu'il n'y a pas le terme B qui

apparaît après leur estimation. En exprimant ceci en prix et non en logarithme du prix, on observe une décote qui va de :

- 18% (intervalle $[17\%, 19\%]$) par an pour le modèle A140 ($\exp(-0.210 \pm 2.02 \times 0.01225)$), à
- 24% (intervalle $[23\%, 25\%]$) par an pour le modèle Omega ($\exp(-0.26390 \pm 2.02 \times 0.01225)$).

où 2.02 est le fractile à 97.5% de la loi de Student à 42 degrés de liberté (loi très proche de la loi normale centrée réduite).

1.4 Un modèle répondant à une question précise

Nous désirons répondre maintenant sur notre petit jeu de données à la questions suivante : est-il vrai, comme il est généralement admis, que les modèles haut de gamme et essence se décotent plus vite que les modèles bas de gamme ou Diesel ?

Pour cela nous avons crée les variables qualitatives :

- `hg` comme haut de gamme caractérise les modèles dont le prix de vente neuf est supérieur à 17000 Euros (prix en 2002).
- `diesel` qui caractérise les modèles diesels.

Nous voulons donc introduire les effets `hg*age` et `diesel*age`. Ici un problème se pose car ces effets définissent des sous-espaces de l'espace engendré par `modele*age`, il ne peuvent pas admettre de test de type III. On est donc obligé de les définir en type I, en prenant soin de mettre tous les autres effets avant. Arbitrairement nous avons choisi de placer `hg*age` avant `diesel*age` et de lancer la suite de commandes :

```
proc glm data=sasuser.argus;
class modele neuf hg diesel;
model log=age modele neuf hg*age diesel*age age*modele/ss1;
run;quit;
```

ce qui donne le résultat suivant :

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age*hg	1	0.00107985	0.00107985	0.57	0.4562

age*diesel	1	0.00241202	0.00241202	1.26	0.2674
age*modele	6	0.03485668	0.00580945	3.04	0.0146

Au vu de ces résultats il est clair que ces hypothèses ne sont pas vérifiées sur nos données.

1.5 Un modèle simple

La dernière étude que nous avons entrepris sur cette base de données a été d'évaluer le pouvoir prédictif du modèle sans hétérogénéité des pentes. Ceci peut être établi par les commandes :

```
proc mixed data=sasuser.argus method =ml;
class modele neuf;
model log=modele neuf age/solution;
run;quit;
```

qui donne les résultats suivants : $R^2 = 0.972$, $AIC \simeq -183.7$, $AICC \simeq -177.2$, $BIC \simeq -158.4$, le coefficient devant `age` vaut environ -0.2220, la décote est d'environ -0.1368, et $\widehat{\sigma} \simeq 0.04869$.

Ce modèle très simple comprend donc par rapport au prix de vente :

- une décote immédiate de 13% car $\exp(-0.1368) \simeq 0.87)$;
- une décote annuelle de 20% car $(\exp(-0.2220) \simeq 0.8)$.

Ce modèle prédit en moyenne avec une erreur relative de 5%, ce qui est comparable aux erreurs des autre modèles nus paraît peut -être le plus séduisant de tous. Pour une estimation plus précise, achetez l'Argus...

2 Etude de cas "Béton" : sélection de modèles et analyse de la covariance

2.1 Présentation de l'expérience

Cette étude de cas reprend de façon parcellaire les résultats d'un travail en collaboration entre le L.S.P. et le L.M.D.C. (Laboratoire des Matériaux et Durabilité des

Constructions) de l'Université Toulouse III, effectué en 2002 (voir la thèse de doctorat de Guillaume Lemaire, 2003).

Le but de l'étude est de déterminer les différents agents de formulation réellement influents sur les propriétés de surface d'un béton brut de décoffrage et d'essayer d'obtenir un modèle permettant d'obtenir des prévisions quant à ces propriétés de surface. Pour cela, une expérience en laboratoire a été effectuée par la constitution de 82 blocs de béton différents, 41 ayant été obtenus à partir d'un coffrage en métal, et 41 à partir d'un coffrage en bois. Dans chacun de ces blocs, on a fait varier le type de ciment utilisé, le type d'adjuvant rajouté,..., et différentes quantités comme la proportion de ciment du mélange, la proportion de granulats,... (plus de détails vont être apportés par la suite). Différentes mesures (quantitatives) des propriétés dites "de surface" ont été effectuées pour chaque bloc ; nous ne nous intéresserons qu'à deux d'entre elles, la luminosité moyenne du béton (obtenue à partir d'un appareil photographique numérique) et la proportion moyenne des bulles apparaissant à la surface du béton.

La démarche retenue est celle d'une modélisation linéaire pouvant associée les différentes variables (potentiellement) explicatives, qu'elles soient qualitatives (5) ou quantitatives (12). Devant le nombre considérable de ces variables en regard du nombre individus (blocs de bétons) disponibles, une sélection de modèle doit d'abord être effectuée. Celle-ci portera d'abord sur les variables quantitatives (à l'aide du critère BIC), puis sur un modèle d'analyse de la variance mélangeant les variables quantitatives retenues et les variables qualitatives.

2.2 Les différentes variables intervenant dans l'étude

Les variables intervenant dans la formulation du béton et potentiellement influentes sur les propriétés de surface sont les suivantes :

Présentation des variables potentiellement explicatives

- cinq variables qualitatives (ou facteurs) liées aux différents matériaux intervenant dans la constitution du béton :

  - `CIM`, le type de ciment employé : `Martres`, `PMES` ou `blanc`.
  - `AJOU`, le type d'ajout associé au béton : `CV`, `FC`, `FS` ou `R` (sans ajout).
  - `ADJU`, le type d'adjuvant employé : `Plastifiant`, `PC`, `PNS`, `PA`, `PMS` ou `PPM`.
  - `COF`, le type de coffrage utilisé : bois (`B`) ou métal (`M`).
  - `BET`, le "type" de béton : classique (`c`) ou spécialisé - auto-plaçant (`s`).

- onze variables quantitatives relatives aux différents matériaux intervenant dans la constitution du béton :

- `Cimval`, quantité de ciment employée.
- `Ajouval`, quantité d'ajout employée.
- `Adjuval`, quantité d'adjuvant employée.
- `granu0315`, la proportion de sable employée de diamètre compris entre 0 et 0.315 mm.
- `granu1`, la proportion de sable employée de diamètre compris entre 0.315 mm et 1 mm.
- `granu4`, la proportion de sable employée de diamètre compris entre 1 mm et 4 mm.
- `granu8`, la proportion de sable employée de diamètre compris entre 4 mm et 8 mm.
- `granu12`, la proportion de sable employée de diamètre compris entre 8 mm et 12 mm.
- `granu20`, la proportion de sable employée de diamètre compris entre 12 mm et 20 mm.
- `GS`, le rapport G/S de granulats par rapport au sable.
- `E-S`, le rapport E/L, eau sur liant.

– une variable quantitative issue d'une mesure sur béton frais :

- `Affaissement`, affaissement du béton classique ou l'étalement du béton autoplaçant.

Présentation des variables d'intérêt et étude descriptive

Ce sont des variables représentatives des propriétés de surface qui motivent cette étude. Elles sont quantitatives et plus précisément nous nous limiterons à :

- la clarté (luminosité) moyenne du béton `Lmoy` à partir d'une photo numérique.

- le bullage du béton, quantifié par `Bulles`, le pourcentage de bulles à la surface du béton.

Ces deux variables sont très faiblement corrélées (corrélation de $\simeq -0.15$.

Corrélations des variables quantitatives

Avant toute chose, précisons qu'il existe une liaison déterministe entre les variables relatives à la proportion des différents granulats et du sable. On a ainsi, par définition :

```
GS= (granu8+granu12+granu20)/(granu0315+granu1+granu4).
```

De plus, l'échantillon ayant été conçu à partir d'un plan d'expériences, il existe également une relation linéaire entre ces 7 variables, `GS, granu0315, granu1, granu4, granu8, granu12, granu20`. Des analyses qui vont suivre, nous devrons parfois prendre en compte cette relation et éliminer arbitrairement une de ces variables pour obtenir un modèle régulier.

Nous avons étudié les différentes variables quantitatives pour déceler leurs éventuelles dépendances. En premier lieu, l'étude des corrélations linéaires donne le résultat suivant :

	Cimval	Ajouval	Adjuval	granu0315	granu1	granu4	granu8	granu12	granu20	E-L	GS
Cimval	1.000	-0.378	0.545	-0.064	-0.318	-0.368	-0.161	0.256	0.107	-0.892	0.249
Ajouval	-0.378	1.000	0.076	0.428	0.132	-0.409	0.627	-0.625	-0.626	0.144	-0.617
Adjuval	0.545	0.076	1.000	0.178	-0.276	-0.479	0.127	-0.135	-0.270	-0.712	-0.249
granu0315	-0.064	0.428	0.178	1.000	-0.554	-0.127	0.576	-0.708	-0.606	0.048	-0.773
granu1	-0.318	0.132	-0.276	-0.554	1.000	-0.392	0.317	-0.163	-0.233	0.418	-0.029
granu4	-0.368	-0.409	-0.479	-0.127	-0.392	1.000	-0.710	0.601	0.719	0.353	0.518
granu8	-0.161	0.627	0.127	0.576	0.317	-0.710	1.000	-0.955	-0.980	0.218	-0.912
granu12	0.256	-0.625	-0.135	-0.708	-0.163	0.601	-0.955	1.000	0.940	-0.282	0.964
granu20	0.107	-0.626	-0.270	-0.606	-0.233	0.719	-0.980	0.940	1.000	-0.110	0.943
E-L	-0.892	0.144	-0.712	0.048	0.418	0.353	0.218	-0.282	-0.110	1.000	-0.214
GS	0.249	-0.617	-0.249	-0.773	-0.029	0.518	-0.912	0.964	0.943	-0.214	1.000
Affaissement	-0.134	0.523	0.234	0.589	0.245	-0.659	0.953	-0.919	-0.969	0.173	-0.924

On peut ainsi observer (comme on pouvait s'y attendre) de fortes corrélations (respectivement positives et négatives) entre respectivement le rapport G/S `GS` et l'affaissement `Affaissement`, avec les variables `granu12` et `granu20` qualifiant les gros granulats. Les variables `Affaissement` et `GS` sont elles-mêmes fortement corrélées négativement. De même, le rapport E/L `E-L` est très lié à la quantité de ciment `Cimval` introduite, ainsi, mais de façon moins nette, qu'à la quantité d'adjuvant `Adjuval`.

Le calcul du VIF (variance inflation factor) des différentes variables quantitatives est plus démonstratif des dépendances linéaires signalées auparavant. On ainsi avec le logiciel R :

```
>vif(lm(bulles~GS+EL+...+granu8+granu12+granu20))}

           GS            EL  Affaissement        Cimval       Ajouval
-6.447540e+15  3.197148e+03  1.749218e+01  4.951419e+01  1.119619e+01
      Adjuval     granu0315        granu1        granu4        granu8
 6.775780e+00 -1.568923e+15 -8.001740e+14 -3.022387e+13 -1.377041e+16
      granu12       granu20
-2.535684e+15 -1.364175e+16
```

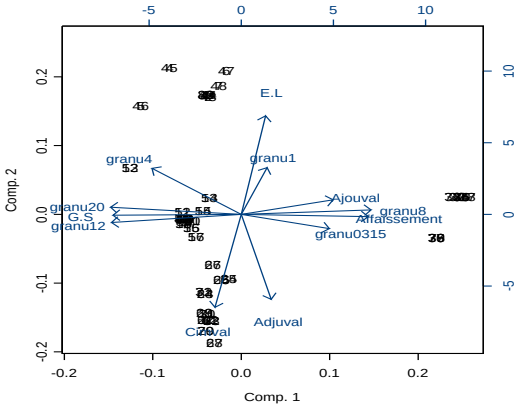
On remarque donc les valeurs très importantes du VIF pour les variables relatives

aux granulats : on retrouve bien le lien linéaire entre ces variables. Par ailleurs, la seule variable semblant un peu moins liée aux autres est `Adjuval`, les autres valeurs du VIF étant très supérieures à 1.

Analyse en composantes principales des variables quantitatives

Pour avoir une idée un peu plus précise de la façon dont les différentes variables explicatives (quantitatives) se comportent, nous avons étudier leur analyse en composantes principales (en abrégé, ACP), ce qui revient à chercher des variables dites principales qui sont des combinaisons linéaires des autres variables et qui concentreraient l'essentiel de "l'information". Les résultats sont les suivants (seules les 4 premières composantes sont prises en compte) :

	Comp. 1	Comp. 2	Comp. 3	Comp. 4
Standard deviation	2.503	1.690	1.292	0.800
Proportion of Variance	0.522	0.238	0.139	0.053
Cumulative Proportion	0.522	0.760	0.899	0.952
Cimval		-0.533	0.117	-0.288
Ajouval	0.276			0.850
Adjuval		-0.487		0.225
granu0315	0.265		-0.555	-0.122
granu1		0.268	0.669	
granu4	-0.269	0.264	-0.427	
granu8	0.391			-0.126
granu12	-0.391			0.114
granu20	-0.394			
E.L		0.565		-0.187
GS	-0.386		0.153	
Affaissement	0.382			-0.229



Les conclusions sont assez proches de celles obtenues par l'étude de la matrice de corrélation. En premier lieu, les 3 premières composantes principales concentrent 90% de l'information (au sens de la variance) de l'ensemble des données quantitatives (pour les 4 premières, c'est 95% de l'information). On peut donc estimer que les premiers 3 axes principaux sont suffisants pour émettre des interprétations sur ces données :

- le premier axe, qui concentre plus de 50% de la variance expliquée, est composé des 5 variables fortement corrélées observées précédemment : `granu20`, `granu12`, `granu8`, `GS` et `Affaissement`, à presque égale proportion. Cette axe peut s'interpréter comme celui des granulats de grande taille.

- le deuxième axe, qui concentre près de 25% de la variance expliquée, est composé des 3 variables corrélées observées précédemment : `E-L`, `Cimval`, et `Adjuval`, ici encore à presque égale proportion. Cet axe peut s'interpréter comme celui du rapport eau sur liant.

- enfin, le troisième axe, d'importance plus faible (15%) est essentiellement celui des variables `granu0315`, `granu1` et `granu4`. On le considérera donc comme celui des granulats de petites tailles (sables).

Enfin, le biplot suivant les deux premiers axes principaux permet non seulement d'observer la part de chacune des variables suivant ces axes, mais également la distribution des individus (82 bétons différents) par rapport à ces axes. On observe ainsi plusieurs "paquets" distincts qui sont pour l'essentiel représentatifs des différents type de bétons. On voit notamment, dans la direction positive du premier axe principal

(donc avec une forte part de granulats de faibles tailles et une grande valeur pour l'affaissement), les bétons auto-plaçants.

2.3 Sélection des variables explicatives

Dans tout ce qui suit nous utiliserons le logiciel R.

On a commencé par constituer un "data.frame", nommé `beton`, comprenant l'ensemble des réalisations des variables explicatives et à expliquer. Ensuite, on crée un autre "data.frame" nommé `beton.exp`, contenant uniquement les 17 variables explicatives. Nous avons choisi de faire de la sélection de modèle pour les variables `Lmoy` et `Bulles` à partir du critère BIC.

Nous allons nous heurter à une difficulté majeure de la sélection de modèle : lorsque le nombre de variables est important, il est impossible de traiter de façon exhaustive tous les modèles possibles car ils sont trop nombreux (dans notre exemple, si l'on ne considère que les modèles additifs, il y a déjà $2^{17} \simeq 131000$ modèles ; si l'on s'intéresse à des modèles avec interactions doubles, on arrive tout de suite à $2^{17*17} \simeq 10^{87}$ modèles possibles !). On essayera donc, en utilisant également les présupposés du problème traité, d'obtenir progressivement un bon modèle, sans cependant être certain d'avoir obtenu le meilleur.

Pour commencer, nous ne regardons que des modèles additifs (sans interaction), ce qui peut se faire par les commandes (en ce qui concerne la variable `Lmoy`) :

```
beton.exp=beton[,2:18]
library(MASS)
Lmoy.lm=lm(beton$Lmoy~.,data=beton.exp)
Lmoy.bic1=stepAIC(Lmoy.lm,k=log(82))
```

La procédure de recherche itérative de modèles par le critère BIC n'a fait qu'éliminer 4 variables (notons que la variable `GS` est éliminée automatiquement car colinéaire aux différentes variables `granu`) ; il en reste 13, ce qui n'est pas très satisfaisant au premier abord. En fait **cela est dû aux algorithmes de `stepAIC` qui arrête d'enlever des termes du modèle dès que l'on a trouvé un minimum local. On s'aperçoit dans ce cas comme dans d'autres que l'on a intérêt à relancer `stepAIC` en forçant la suppression de nouvelles variables.** Après quelques manipulations on retient deux modèles qui ont des BIC presque égaux :

- Un modèle avec 6 variables, dont le R^2 vaut $\simeq 0.77$:

```
Step:  AIC= 187.17
 beton$Lmoy ~ Bet + Cim + Cimval + Ajou + EL + Cof

         Df Sum of Sq     RSS     AIC
<none>                  469.60  187.17
- EL      1     27.56  497.15  187.44
- Cof     1     28.24  497.83  187.55
- Cimval  1     42.89  512.48  189.93
- Bet     1     73.43  543.03  194.68
- Ajou    3    164.89  634.49  198.63
- Cim     2   1319.03 1788.63  288.02
```

- Un modèle avec une seule variable `Cim` dont le R^2 vaut $\simeq 0.67$:

```
Step:  AIC= 187.2
 beton$Lmoy ~ Cim

       Df Sum of Sq     RSS     AIC
<none>                684.29  187.20
- Cim   2   1359.09 2043.38  268.09
```

On peut également étudier les graphes des résidus/valeurs prédites (voir Figure 2.3). Les deux graphes, et surtout celui du modèle avec la seule variable `Cim`, ne sont pas très satisfaisants en raison d'une apparente hétéroscédasticité. Nous allons essayer d'améliorer ces modèles en utilisant des interactions d'ordre 2 dans des modèles reprenant les 6 variables sélectionnées, ainsi que les deux variables quantitatives complémentaires aux variables `Ajou` et `Bet`, qui sont respectivement les variables `Ajouval` et `Affaissement` :

```
attach(beton.exp)
Lmoy.lm5=lm(beton$Lmoy~(Cim+Cimval+Bet+Ajou+Cof+Ajouval+Affaissement+EL)^2)
Lmoy.bic5=stepAIC(Lmoy.lm5,k=log(82))
Anova(Lmoy.bic5,type="III",contrasts=list(Cim=contr.sum,Cimval=contr.sum,
      Bet=contr.sum,Ajou=contr.sum,Cof=contr.sum,Ajouval=contr.sum,
                                    Affaissement=contr.sum,EL=contr.sum))
```

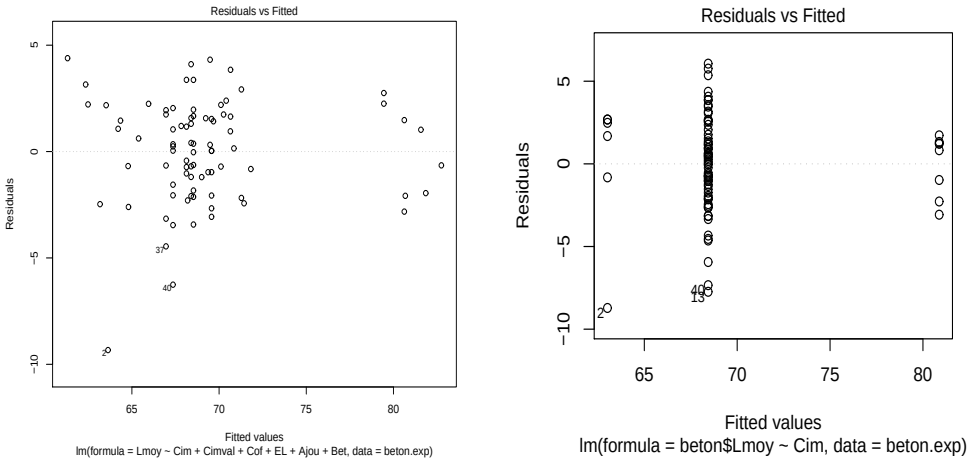


FIGURE 15.1 – Graphique des résidus en fonction des valeurs prédites pour les modèles avec 6 variables retenues (à gauche) et avec seulement la variable `Cim` (à droite)

(**Rappelons que dans la nomenclature de R, le terme croisé** $Cim * Cimval$ **contient l'interaction** $Cim : Cimval$ **ainsi que les facteurs principaux** Cim **et** $Cimval$**)**. Finalement, on aboutit au modèle $Lmoy \sim Cim * Cimval + Bet + Cof + Ajou + EL$ pour lequel le critère BIC vaut $\simeq 172.81$ (c'est-à-dire beaucoup moins que pour les deux modèles précédents sans interaction) et $R^2 \simeq 0.83$, les tentatives pour enlever d'autres variables dans la procédure de minimisation du critère BIC ne s'avérant pas fructueuses. Notons que par la commande `stepAIC(lm(beton$Lmoy~.^2,data=beton.exp))`, on peut sélectionner un modèle par le critère BIC à partir du modèle contenant toutes les interactions doubles possibles, mais le résultat de cette procédure, en prenant un temps de calcul beaucoup plus important, a fini par mener au même modèle que celui auquel nous sommes arrivés en étant parti d'un modèle contenant les variables sélectionnée à partir de modèles additifs.

Il nous reste à examiner la possibilité d'interactions d'ordres supérieurs (on ne peut monter trop haut dans les ordres possibles, sachant que notre échantillon est de taille 82 et que les variables qualitatives peuvent posséder jusqu'à 6 modalités (c'est le cas de `Ajou`). Mais cet examen nous conduit à nouveau au même modèle $Lmoy \sim Cim * Cimval + Bet + Cof + Ajou + EL$.

De tout ceci, on retient que les variables semblant réellement expliquer la variable `Lmoy` sont : `Cim*Cimval, Bet, Cof, Ajou` et `EL`, la variable qualitative `Cim` étant clairement la plus importante.

On applique également la même méthodologie à la variable d'intérêt `Bulles`. On obtient à l'aide du critère BIC et dans le cadre d'un modèle sans interaction, un modèle (tel que $BIC \simeq -159.5$ et $R^2 \simeq 0.47$) contenant 4 variables `Cof, Bet, Affaissement` et `Adjuval`, modèle validé par des tests de Fisher (avec 5 coefficients estimés). En étudiant maintenant des modèles avec interactions, on en arrive au modèle final (tel que $BIC \simeq -165.03$ et $R^2 \simeq 0.66$), $Bulles \sim Bet * Affaissement + Adjuval * Cof + Adju$, modèle (avec 12 coefficients estimés) validé par des tests de Fisher. Cependant, si il est assez logique que les variables `Bet` et `Affaissement` soient associées (la mesure de l'affaissement (de l'étalement) du béton est liée avec le fait que celui-ci soit classique ou auto-plaçant), on est un peu surpris de voir la variable `Adjuval` associée à la variable `Cof` et non avec la variable `Adju`.

Voici un tableau récapitulatif des variables explicatives sélectionnées, les modèles retenus laissant pour l'instant à désirer :

Variable à expliquer	R2	BIC	Variables explicatives	Nombre de coef
`Lmoy`	0.83	172.81	`Cim*Cimval, Bet, Ajou, Cof` et `EL`	12
`Bulles`	0.66	−165.01	`Cof*Adjuval`, `Bet*Affaissement` et `Adju`	11

2.4 Amélioration des modèles

Reprenons l'étude de la variable `Lmoy`. La double figure 2.3 nous montre qu'un modèle linéaire simple n'est peut-être pas la meilleure solution en vue de prédiction. En effet, il apparaît que si l'on considère le modèle linéaire $Lmoy \sim Cim$, l'écart-type des résidus ne semble pas indépendant des valeurs prédites (c'est-à-dire à peu près constant) : $\widehat{\sigma}(\texttt{blanc}) \simeq 1.85$, $\widehat{\sigma}(\texttt{Martres}) \simeq 2.89$ et $\widehat{\sigma}(\texttt{PMES}) \simeq 4.47$. Si on couple ces valeurs avec les valeurs prédites, on obtient une courbe clairement décroissante. Il peut donc être intéressant de faire un changement de variable pour la variable `Bulles`. Après diverses tentatives, on trouve que le changement de variable $\texttt{Lmoy} \longrightarrow (\texttt{Lmoy})^4$ permet d'avoir un écart-type à peu près constant. On s'aperçoit également de l'intérêt de ce modèle $(Lmoy)^4 \sim Cim$ lorsque l'on considère le coefficient de détermination $R^2 \simeq 0.76$ (à comparer avec $R^2 \simeq 0.67$ pour le modèle initial $Lmoy \sim Cim$. Ce changement de variable a donc été fructueux (notons cependant qu'il n'est plus possible d'utiliser le critère BIC pour effectuer une comparaison ces deux modèles). Il reste maintenant à enrichir le modèle des autres variables sélectionnées pour améliorer encore la prédiction.

On peut à nouveau utiliser le critère BIC pour sélectionner le modèle linéaire vérifié

par $(\texttt{Lmoy})^4$. On commence par reprendre les variables précédemment choisies, ce qui peut se faire par les commandes :

```
Lmoy4.lm=lm((beton$Lmoy)^4~(Cim+Cimval+Bet+Ajou+Cof+Ajouval+Affaissement)^2)
Lmoy4.bic=stepAIC(Lmoy4.lm,k=log(82))
Anova(lm(I(beton$Lmoy)^4~Cim*Cimval+Affaissement+Ajou),type="III",
      contrasts=list(Cim=contr.sum,Ajou=contr.sum))
```

Le modèle obtenu (par minimisation répétée du critère BIC) est différent du précédent : les variables `EL`, `Cof` et `Bet` semblent ne plus devoir intervenir, mais la variable `Affaissement` jouerait un rôle cependant, la différence du critère BIC entre les deux modèles est très faible). On retient cependant le modèle $Lmoy \sim Cim * Cimval + Ajou + Affaissement$ qui ne contient que 4 variables (et 8 coefficients estimés), a donc une valeur de BIC apparemment minimale ("apparemment", car, encore une fois, on n'a pu tester de façon exhaustive tous les modèles) et une valeur $R^2 \simeq 0.855$ très proche de celle obtenue avec le modèle $Lmoy \sim Cim*Cimval+Ajou+Cof+Bet+EL$ ($R^2 \simeq 0.868$). Pour terminer, on peut essayer de modéliser $(\texttt{Lmoy})^4$ par des fonctions des variables quantitatives choisies (c'est-à-dire `Cimval` et `Affaissement`). Le critère BIC va nous aider à déterminer une telle fonction. On peut ainsi taper les commandes suivantes :

```
Lmoy4.lm=lm((beton$Lmoy)^4~(Cim+Cimval+I(Cimval^2)+I(Cimval^0.5)+Ajou+
    Affaissement+I(Affaissement^0.5)+I(Affaissement^0.5))^2)
Lmoy4.bic=stepAIC(Lmoy4.lm,k=log(82))
```

On aboutit à un modèle privilégiant la variable $(\texttt{Affaissement})^{0.5}$ à `Affaissement`, mais en revanche, `Cimval` semble finalement plus significative que $\texttt{Cimval}^2$ ou que $\texttt{Cimval}^{1/2}$: on en restera là quant à la variable `Cimval`. On continue de chercher une fonction "moins croissante" que la fonction identité affectée à la variable `Affaissement` :

```
Lmoy4.lm=lm((beton$Lmoy)^4~Cim*Cimval+Ajou+I(Affaissement^0.5)+I(1/Affaissement)
    +I(log(Affaissement)))
Lmoy4.bic=stepAIC(Lmoy4.lm,k=log(82))
```

Après plusieurs tentatives, on sélectionne finalement le modèle : $Lmoy \sim Cim * Cimval + Ajou + 1/Affaissement$, pour lequel $BIC \simeq 2478.43$ (contre $BIC \simeq 2480.65$ avec `Affaissement`) et $R^2 \simeq 0.859$ (contre $R^2 \simeq 0.855$ avec `Affaissement`), modèle validé par des tests de Fisher. Les deux graphiques de la figure 2.4 confirment le choix de ce modèle.

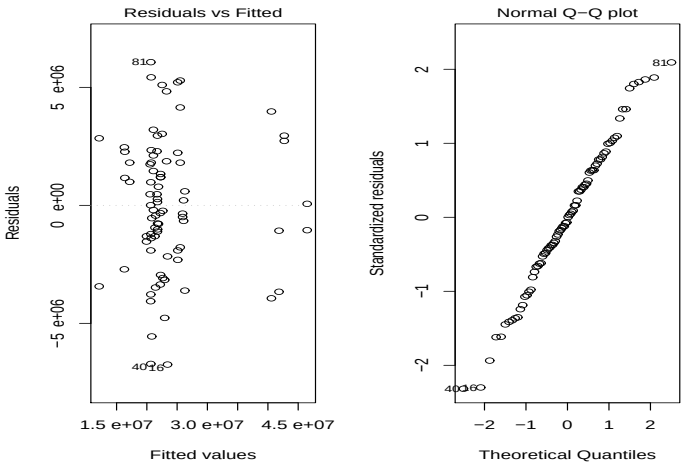


FIGURE 15.2 – Graphiques de contrôle du modèle $Lmoy \sim Cim * Cimval + Ajou + 1/Affaissement$

De la même manière, on essaye d'améliorer le modèle pour la variable `Bulles`. Pour commencer, on examine les graphiques de contrôle (Figure 2.4) du modèle linéaire choisi précédemment. On constate une forme en entonnoir du graphe des résidus en fonction des valeurs prédites, et un QQ-plot pas assez linéaire. On va remédier à ces défauts en effectuant le changement de variable `Bulles` $\longrightarrow$ $(\texttt{Bulles})^{1/2}$ (obtenu graphiquement et par l'observation du R^2). Une nouvelle sélection d'un modèle avec interaction par le critère BIC (en reprenant cependant les variables déjà choisies) amène au même modèle : $(Bulles)^{1/2} \sim Bet * Affaissement + Adjuval * Cof + Adju$.

Enfin, comme précédemment, on essaye de remplacer les deux variables quantitatives présentes par des fonctions de ces variables aboutit au modèle final : $(Bulles)^{1/2} \sim Bet + (Affaissement)^2 + Adjuval + Cof + Adju$, qui est à nouveau un modèle additif. La figure 2.4 présente les graphes de contrôle de ce modèle. On apprécie les qualités de ce nouveau modèle par rapport au précédent sur ces graphiques. Les tests de Fisher valident ce modèle seul le coefficient $R^2 \simeq 0.66$ est un petit peu moins bon que précédemment ($R^2 \simeq 0.68$).

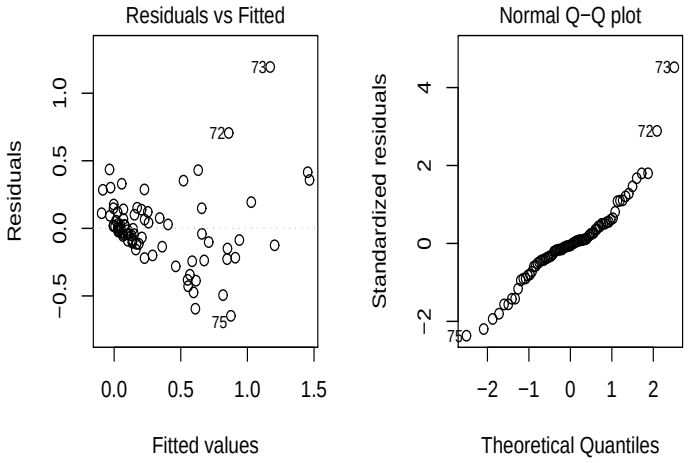


FIGURE 15.3 – Graphiques de contrôle du modèle $Bulles \sim Bet * Affaissement + Adjuval * Cof + Adju$

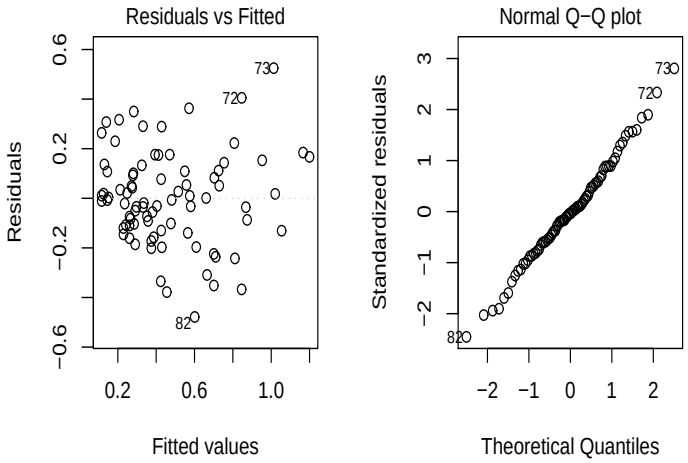


FIGURE 15.4 – Graphiques de contrôle du modèle $(Bulles)^{1/2} \sim Bet + (Affaissement)^2 + Adjuval + Cof + Adju$

2.5 Conclusions

En conclusion et en vue de prédictions, on a obtenu les modèles :

Modèle	R2	Nombre de coefficients à est
$(\texttt{Lmoy})^4 \sim \texttt{Cim} * \texttt{Cimval} + \texttt{Ajou} + (1/\texttt{Affaissement})$	0.86	10
$(\texttt{Bulles})^{1/2} \sim \texttt{Bet} + \texttt{Adju} + \texttt{Cof} + \texttt{Adjuval} + (\texttt{Affaissement})^2$	0.66	10

Voici également les coefficients estimés de ces deux modèles :

- Pour la luminosité moyenne $(\texttt{Lmoy})^4$:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
Cimblanc           53050105    7184942   7.384 2.16e-10 ***
CimMartres         30486749    2616994  11.650  < 2e-16 ***
CimPMES            -4974907    7271215  -0.684  0.49605
Cimval               -34883      18673  -1.868  0.06582 .
AjouFC              6745090    1537558   4.387 3.86e-05 ***
AjouFS              4317747    2277251   1.896  0.06197 .
AjouR               6155641    1062179   5.795 1.67e-07 ***
I(1/Affaissement) -56115642   16436568  -3.414  0.00105 **
CimMartres:Cimval      7672      20126   0.381  0.70419
CimPMES:Cimval        84524      26816   3.152  0.00236 **
```

- Pour le nombre de bulles $(\texttt{Bulles})^{1/2}$:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
Betc                 5.894e-01  1.085e-01   5.432 7.19e-07 ***
Bets                 1.953e+00  2.536e-01   7.703 5.48e-11 ***
AdjuPC              -3.943e-01  1.074e-01  -3.670 0.000462 ***
AdjuPlastifiant     -4.141e-01  1.154e-01  -3.587 0.000607 ***
AdjuPMS              6.904e-02  1.462e-01   0.472 0.638194
AdjuPNS             -5.160e-02  1.440e-01  -0.358 0.721173
AdjuPPM             -2.774e-01  1.440e-01  -1.927 0.057967 .
CofM                 1.454e-01  4.477e-02   3.247 0.001773 **
Adjuval              3.733e-02  7.554e-03   4.941 4.88e-06 ***
I((Affaissement)^2) -2.810e-04  5.397e-05  -5.206 1.75e-06 ***
```

Commentaire final quant aux résultats de l'étude : si vous désirez préparer vous même votre béton en essayant d'avoir un "beau" béton ("beau" étant entendu dans le sens de l'opinion la plus commune, c'est-à-dire un béton bien clair possédant le moins de bulles possible), il vous faudra mettre dans votre mélange initial un ciment très clair (de type `blanc`) plutôt en faible quantité, un ajout de type `FC` (ou pas d'ajout), suffisamment d'eau (pour augmenter l'affaissement), rajouter un adjuvant de type `PC` ou `Plastifiant` en faible quantité, mettre suffisamment de "gros" granulats pour obtenir un béton classique (de type `c`), et laisser mijoter le tout dans un coffrage en bois...

3 Etude de cas "Cola" : Plans d'expériences

3.1 Présentation de l'expérience

Nous allons présenter une application de la méthode des plans d'expériences sur un sujet de tous les jours : la comparaison de certaines boisson au Cola. Nous décrirons la construction de l'expérience en utilisant la structure factorielle du problème, puis nous décrirons l'analyse des résultats qui amène à quelques calculs de probabilités qui ne sont pas classiques.

Une expérience menée à l'Université Paul Sabatier par deux étudiants : Sylvan Rey et Julie Trappier [51] a pour but de comparer trois marques de boisson au Cola. Deux sont des marques très connues, nous les noterons `C` et `P`, la troisième est un produit dit "générique" : `G`.
La comparaison utilise la technique des plans d'expériences, mais de manière atypique puisque la réponse est qualitative. En effet, pour maximiser la puissance de l'expérience, les deux étudiants utilisé la technique des "comparaisons ternaires" dite encore "solo-duo" : on présente à un dégustateur 3 verres identiques extérieurement dont deux sont remplis d'un même produit alors que le troisième contient un autre produit : "l'intrus". Après avoir goûté les 3 verres, le dégustateur doit :

- désigner l'intrus ;
- dire quel produit il préfère.

La méthodologie des plans d'expériences doit s'appliquer car la question n'est pas si simple qu'il paraît. En effet, il existe des produits

- sucrés et des produits dit "lights" aux édulcorants ;

- sous emballage plastique et sous emballage métallique ;
- parfumés (modalité `citron`) ou non parfumés (modalité `nature`) ;
- enfin, la dégustation peut se faire à température ambiante (modalité `ambiant`) ou à température du réfrigérateur (modalité `frais`).

Pour des question de budget, puisqu'il s'agit d'une expérience de type "amateur", nous n'avons pas pu aborder, en dehors de la comparaison plastique-métal, un point pourtant très important comme l'analyse de la variabilité inter-lots. Pour supprimer au contraire cette variabilité, tous les produits correspondant à une même dénomination ont été achetés le même jour dans un supermarché et ce à l'intérieur du même emballage plastique réunissant plusieurs bouteilles ou canettes métalliques.

Notre budget a ainsi permis de faire 48 comparaisons ternaires faites par 16 dégustateurs qui étaient des étudiants volontaires. La dégustation s'est faite à 11 Heures du matin.

En fait l'expérience est constituée de 5 sous-expériences ayant des buts différents.

i. La comparaison des deux marques `C` et `P`. C'est ce que nous appellerons l'expérience principale. En ce qui concerne les facteurs

   - `sucre` (modalités `light` ou `sucre`) ;
   - `temperature` (modalités `ambiant` ou `frais`) ;
   - `emballage` (modalités `plastique` ou `metal`) ;
   - `plateau` : on choisit le produit qui est présenté deux fois, et celui qui est présenté une fois.

   Nous avons ainsi réalisé un plan factoriel complet qui demande 16 unités expérimentales, c'est-à-dire 16 plateaux de 3verres.

ii. La comparaison nature-parfumé. Elle comprend 8 unités avec structure de plan factoriel complet sur les facteurs : `produit`, `temperature` et `plateau`.

iii. La comparaison marque - générique : 8 unités basées sur les facteurs `produit`, `plateau` et `temperature`.

iv. La comparaison light-sucré, 8 unités. Il s'agit d'un plan fractionnaire 2^{4-1} basé sur les 4 facteurs `produit`, `plateau`, `temperature` et `emballage`.

v. La comparaison plastique-métal. Il s'agit d'un plan fractionnaire 2^{4-1} basé sur les 4 facteurs `produit`, `plateau`, `temperature` et `sucre`.

3.2 Construction de l'expérience

Chacun des 16 dégustateurs physiques tire au hasard un numéro de dégustateur pour l'expérience principale, ainsi que deux numéros dans deux parmi les quatre expériences secondaires. Le préparateur verse les liquides dans les verres. Le serveur qui ne connaît pas le contenu du plateau l'amène et appelle le dégustateur concerné. Le dégustateur teste alors les trois verres et coche sur sa fiche l'intrus puis le produit qu'il a préféré. Chaque dégustateur a donc à remplir trois grilles du type suivant :

DEGUSTATEUR N°

Comparaison i :

	A	B	C
Intrus			
Préféré			

Nous décrivons maintenant la construction par SAS de l'expérience principale. Comme chaque dégustateur se voit attribuer un numéro au hasard, cela suffirait comme randomisation si l'on n'avait pas à choisir le numéro de l'intrus : facteur `numero` qui sera codé A, B et C dans nos résultats (mais 1, 2 et 3 par commodité au niveau SAS). Voici donc le source correspondant à la création du plan. Il utilise la procédure `proc plan`.

```
proc plan;
factors emballage=2 sucre=2 temperature=2 plateau=2;
treatment numero=3;
output out=sasuser.cola1 emballage sucre temperature plateau;
proc print; run;
```

La procédure crée donc une expérience comprenant 16 unités correspondant aux 16 combinaisons (avec des niveaux aléatoires) des 4 facteurs déclarés comme `factors`. Sur cette structure on tire au hasard le facteur `numero` à trois niveaux. A la fin tout est sauvé sur la table-data `sasuser.cola1`.

3.3 Les résultats des expériences

Expérience principale

unité	plateau	temperature	sucre	emballage	intrus	résultat	préféré
1	2 C - 1 P	frais	sucre	metal	A	juste	C
2	1 C - 2 P	frais	sucre	metal	B	faux	
3	2 C - 1 P	ambiant	sucre	metal	B	juste	C
4	1 C - 2 P	ambiant	sucre	metal	C	faux	
5	2 C - 1 P	ambiant	light	metal	A	faux	
6	1 C - 2 P	ambiant	light	metal	C	juste	C
7	1 C - 2 P	frais	light	metal	B	juste	P
8	2 C - 1 P	frais	light	metal	A	faux	
9	1 C - 2 P	frais	sucre	plastique	C	faux	
10	2 C - 1 P	frais	sucre	plastique	C	faux	
11	2 C - 1 P	ambiant	sucre	plastique	A	juste	C
12	1 C - 2 P	ambiant	sucre	plastique	B	juste	C
13	2 C - 1 P	frais	light	plastique	B	faux	
14	1 C - 2 P	frais	light	plastique	A	faux	
15	2 C - 1 P	ambiant	light	plastique	B	juste	C
16	1 C - 2 P	ambiant	light	plastique	A	faux	

Il y a eu en tout 7 bonnes réponses sur 16 pour désigner l'"intrus".

Comparaison nature-parfumé

unité	plateau	temperature	produit	intrus	résultat	préféré
17	2 citron - 1 nature	ambiant	C	A	faux	
18	1 citron - 2 nature	ambiant	C	C	juste	citron
19	1 citron - 2 nature	frais	C	B	juste	nature
20	2 citron - 1 nature	frais	C	A	juste	citron
21	1 citron - 2 nature	frais	P	B	juste	nature
22	2 citron - 1 nature	frais	P	B	juste	citron
23	2 citron - 1 nature	ambiant	P	A	faux	
24	1 citron - 2 nature	ambiant	P	A	juste	citron

Il y a eu en tout 6 bonnes réponses sur 8 pour désigner l'"intrus".

Comparaison marque-générique

unité	plateau	temperature	intrus	résultat	préféré
25	2 G - 1 P	frais	B	vrai	générique
26	1 G - 2 P	frais	C	vrai	P
27	2 G- 1 P	ambiant	A	vrai	P
28	1 G - 2 P	ambiant	A	faux	
29	1 G - 2 C	frais	C	vrai	C
30	2 G - 1 C	frais	C	faux	
31	2 G- 1 C	ambiant	B	vrai	générique
32	1 G - 2 C	ambiant	C	faux	

Il y a eu en tout 5 bonnes réponses sur 8 pour désigner l'"intrus".

Comparaison light-sucré

exp	plateau	produit	temperature	emballage	intrus	résultat	préféré
33	2 sucre - 1 light	P	frais	plastique	B	faux	
34	1 sucre - 2 light	P	ambiant	plastique	C	vrai	light
35	2 sucre - 1 light	P	frais	metal	C	faux	
36	1 sucre - 2 light	P	ambiant	metal	B	faux	
37	1 sucre - 2 light	C	frais	plastique	C	vrai	sucre
38	2 sucre - 1 light	C	ambiant	plastique	C	vrai	sucre
39	2 sucre - 1 light	C	frais	metal	C	vrai	sucre
40	1 sucre - 2 light	C	ambiant	metal	C	vrai	sucre

Il y a eu en tout 5 bonnes réponses sur 8 pour désigner l'"intrus".

Comparaison plastique-métal

exp	plateaux	produit	temperature	sucre	intrus	résultat	préféré
41	2 plastique - 1 metal	P	ambiant	light	B	faux	
42	1 plastique - 2 metal	P	frais	light	C	faux	
43	2 plastique - 1 metal	P	ambiant	sucre	B	vrai	plastique
44	2 plastique - 1 metal	P	frais	sucre	C	faux	
45	2 plastique - 1 metal	C	frais	light	B	vrai	métal
46	2 plastique - 1 metal	C	ambiant	light	B	vrai	métal
47	2 plastique - 1 metal	C	frais	sucre	A	vrai	plastique
48	2 plastique - 1 metal	C	ambiant	sucre	A	vrai	métal

Il y a eu en tout 5 bonnes réponses sur 8 pour désigner l'"intrus".

3.4 Analyse des résultats

Première analyse

Dans un premier temps nous n'interprétons pas les préférences et considérons uniquement le nombre de bonnes réponses. Nous sommes amenés à réaliser un test unilatéral. En effet, dans l'hypothèse d'absence d'effet, la probabilité de bonne réponse vaut 1/3 et il semble raisonnable de ne considérer que des alternatives pour lesquelles cette probabilité est supérieure. Le tableau ci-dessous donne le niveau exact du test en fonction du nombre d'essais et de la zone de rejet.

Nb d'essais	zone de rejet	niveau exact
8	≥ 5	0.0879
8	≥ 6	0.0197
16	≥ 9	0.05
24	≥ 12	0.0677
24	≥ 13	0.0284
48	≥ 21	0.086
48	≥ 22	0.0485
48	≥ 23	0.0254

Seulement deux expériences ou sous-expériences sont significatives :

- la comparaison citron-nature (P-Value =0.02) ;
- la comparaison sucré-light sur le produit C : 4 bonnes réponses (P-Value =1/81= 0.012).

Par ailleurs les comparaisons : marque-générique, light-sucré, plastique-métal montrent une légère présomption de significativité (P-Value =0.088). En revanche, le nombre global de bonnes réponses (28) est clairement significatif. Les 15 bonnes réponses à température ambiante comme les 13 à température fraîche le sont également. Le score moyen, d'environ 58%, significativement différent de 1/3, est à rapprocher des scores de comparaisons ternaires sur les vins cités dans la littérature (Morrot et Brochet [48]), avec un taux de découverte de l'intrus qui est de 62%.

En conclusion, dans l'ensemble des produits comparés, il existe clairement des produits différents. Il est par contre très difficile de dire quels produits diffèrent puisque (si on met de coté light-sucré pour le produit `C`) seule la comparaison extrême nature-parfumé est significative.

Si on prend en compte maintenant les préférences, trois remarques s'imposent :

- Pour la comparaison light-sucré sur le produit `C`, la préférence va toujours au produit sucré. Sous l'hypothèse nulle, la probabilité de trouver 4 fois l'intrus et de toujours préférer le même produit est de :

$$(1/3)^4(1/2)^3 = 1.54..10^{-3}.$$

  Cela renforce donc considérablement la significativité de l'effet. Ce phénomène peut être dû également à une variabilité inter-lot, variabilité que nous n'avons pu quantifier faute de moyens.

- Sur la plupart des expériences, les préférences sont équilibrées et ne peuvent être utilisée pour renforcer la significativité. La comparaison nature-parfumé indique que les consommateurs font une différence entre les produits mais que leur préférence semble aller de manière équilibrée à un produit comme à l'autre.

- Il reste le cas de l'expérience principale où, d'une part, les 7 réponses justes et, d'autre part, la répartition des préférence 6 à 1, donnent séparément de légère présomption sans être significatives. Comment combiner les deux informations ? C'est ce que nous allons aborder au paragraphe suivant.

Combinaison des informations

Une première idée serait de réaliser un test du χ^2 (test d'adéquation). En effet, le résultat de l'expérience peut être :

- l'intrus n'est pas trouvé avec probabilité p_1 ;
- l'intrus est trouvé et on préfère le produit C avec probabilité p_2 ;
- l'intrus est trouvé et on préfère le produit P avec probabilité p_3.

On désirera alors tester l'hypothèse nulle

$$p_1 = 2/3 \ , \quad p_2 = 1/6 \ , \quad p_3 = 1/6. \tag{15.1}$$

Dans notre exemple, on dispose de $n = 16$ observations et des effectifs $Y_1 = 9$, $Y_2 = 6$ et $Y_3 = 1$. La statistique du test du χ^2 associé vaut environ 5.4. La valeur critique asymptotique à 5% vaut environ 6.0, mais elle est très approximative compte tenu du faible effectif. Il faudrait donc calculer une valeur critique non asymptotique, mais nous ne l'avons pas fait car le test du χ^2 n'exploite pas l'information que l'on se place dans l'alternative $p_1 < 2/3$. Comme le test du χ^2 est proche du test du rapport de

vraisemblance, nous avons donc préféré effectuer un test du rapport de vraisemblance entre l'hypothèse nulle H_0 donnée par (15.1) et l'hypothèse générale H_1 donnée par $p_1 < 2/3$.

Il est facile de vérifier que la log-vraisemblance sous H_0 vaut

$$LV_0 = Y_1 \cdot \log(2/3) + Y_2 \cdot \log(1/6) + Y_3 \cdot \log(1/6).$$

(on a négligé les coefficients multinomiaux qui sont constants). Dans le cas où $n = 16$, les estimateurs du maximum de vraisemblance sous H_1 sont définis par :

- si $Y_1 \leq 10$: $\widehat{p_i} = n_i/n$, pour $i = 1, 2, 3$;
- si $Y_1 \geq 11$: $\widehat{p_1} = 2/3$, $\widehat{p_2} = \dfrac{Y_2}{3(Y_2 + Y_3)}$ et $\widehat{p_3} = \dfrac{Y_3}{3(Y_2 + Y_3)}$,

et la log-vraisemblance sous H_1 vaut

$$LV_0 = Y_1 \cdot \log(\widehat{p_1}) + Y_2 \cdot \log(\widehat{p_2}) + Y_3 \cdot \log(\widehat{p_3}).$$

Dans le cas de nos observations, le calcul montre que la statistique du rapport de vraisemblance vaut : $d \simeq -13.8358 - (-16.1915) \simeq 2.3557$.

Le programme en R donné en annexe permet de simuler un échantillon de taille n de la distribution d'échantillonnage sous l'hypothèse nulle, ce dans le cas de 16 observations. Il est donc facile d'en déduire la P-value associée à la valeur 2.3557.

Sur 10^7 simulations, on trouve que la probabilité pour que la statistique de test dépasse 2.3557 est d'environ 0.070. Cela donne une légère présomption mais ce n'est pas significatif. Le mystère reste donc entier...

3.5 Conclusion

Notre étude reste, et c'est peut-être mieux ainsi, un exemple d'école. Il conviendrait, si l'on disposait d'un budget plus conséquent :

i. d'étudier la variabilité inter-lot qui n'est approchée ici que par la comparaison métal-plastique ;

ii. de faire des comparaisons avec un effectif plus important.

Faisons un calcul d'effectif pour obtenir un niveau de 5% et une puissance de 90% sous l'alternative dans laquelle la probabilité de bonne réponse est de 60% (ce qui

correspond grosso-modo au score moyen). En utilisant l'approximation gaussienne de la loi binomiale, pour un test unilatéral à 5% sur un effectif de n individus la valeur critique pour le nombre de succès est

$$S = n/3 + 1.65 \cdot \sqrt{\frac{2n}{9}}.$$

Si on veut une puissance de 90% dans le cas d'une probabilité de 0.6, il faut que

$$S = 0.6 \cdot n - 1.28 \cdot \sqrt{n \times 0.4 \times 0.6}$$

(car -1.28 est le fractile associé à la probabilité 0.10 de la loi normale centrée réduite. En résolvant on trouve $n \geq 36$. Pour effectuer les 5 comparaisons de l'expérience ci-dessus ainsi que la comparaison inter-lot, il faudrait donc environ $6 \times 36 = 216$ unités. Cela demande un vrai budget et une vraie organisation...

Notons enfin que la partie de la section 3.4 qui concerne la combinaison des informations propose une méthodologie pour l'analyse statistique des comparaisons ternaires.

3.6 Programme

On commence par prolonger la fonction $x \cdot \log(x)$ en zéro en créant la fonction :

```
xlog=function(x) {if (x>0) {x*log(x)} else 0}
```

puis la simulation se fait par :

```
cola=function(n)
{t=rep(0,n)
for (i in 1:n) {
    Y=rmultinom(c(1,2,3),16,c(2/3,1/6,1/6)) %tirage de la multinomiale
    lv0=Y[1]*log(2/3)+Y[2]*log(1/6)+Y[3]*log(1/6) % estimateur du MV
    if (Y[1]>= 11) {p1=2/3
        if ((Y[3]+Y[2])==0)
            {p2=1/6; p3=1/6}
        else {p2=Y[2]/(3*(Y[3]+Y[2])); p3=Y[3]/(3*(Y[3]+Y[2]))}
        lv1=Y[1]*log(2/3)+3*(Y[3]+Y[2])*(xlog(p2)+xlog(p3))}
    else {p1=Y[1]/16; p2=Y[2]/16; p3=Y[3]/16;
        lv1=16*(xlog(p1)+xlog(p2)+xlog(p3))}
    t[i]=lv1-lv0}
    f=hist(t,br=c(-1,2.3557,100),plot=FALSE)$count
    cola=f[2]/n
}
```

Annexe A

Rappels de Probabilités

Nous rappelons dans cette partie quelques propriétés de la théorie des probabilités qui sont utilisées dans les chapitres précédents. Le but n'est pas absolument pas de fournir les bases du calcul des probabilités (bases qui sont nécessaires à une bonne compréhension de la statistique inférentielle ; une très bonne et assez complète introduction au calcul des probabilités peut se trouver dans le livre Barbe et Ledoux [8]), mais d'exposer quelques résultats qui faciliteront la lecture. Ainsi, seule une démonstration, d'un théorème absolument central pour étudier le modèle linéaire gaussien sera proposée ici.

Hypothèses initiales

Dans tout ce qui suit, on se place sur $(\Omega, \mathcal{A}, P)$ un espace de probabilité, c'est-à-dire un ensemble Ω de résultats possibles d'une expérience aléatoire, un ensemble $\mathcal{A}$ de sous-ensembles de Ω (appelé encore tribu des événements aléatoires possibles issus de l'expérience) et une mesure (ou loi) de probabilité pour les événements de $\mathcal{A}$ (Ω et $\mathcal{A}$ ne seront en général jamais précisés). On appellera X, Y ou Z des variables (ou vecteurs) aléatoires sur $(\Omega, \mathcal{A}, P)$, c'est-à-dire des fonctions mesurables à valeurs dans $\mathbb{R}$ (ou $\mathbb{R}^d$). On suppose par ailleurs que dans les propriétés suivantes, les différents moments (espérance et variance) de ces variables (vecteurs) existent.

1 Règles opératoires du calcul de l'espérance et de la variance d'une variable aléatoire

L'espérance donne la position moyenne théorique d'une variable aléatoire. C'est un opérateur linéaire (de l'espace vectoriel constitué par l'ensemble des variables aléatoires

dans $\mathbb{R}$). Ainsi, si α et β sont des constantes réelles , alors :

$$\mathbb{E}(\alpha \cdot X + \beta \cdot Y) = \alpha \cdot \mathbb{E}(X) + \beta \cdot \mathbb{E}(Y),$$

quelles que soient les relations entre les variables aléatoires X et Y.

La variance mesure l'écart quadratique (théorique) d'une variable aléatoire à son espérance. Elle se définit à partir de l'espérance de la façon suivante :

$$\mathrm{Var}\, X = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Elle est quadratique et invariante par addition d'une constante :

$$\mathrm{Var}(\alpha \cdot X + \beta) = \alpha^2 \cdot \mathrm{Var}(X).$$

La variance de la somme de variables aléatoires fait intervenir la covariance entre ces variables :

$$\mathrm{Var}(\alpha \cdot X + \beta \cdot Y) = \alpha^2 \cdot \mathrm{Var}(X) + \beta^2 \cdot \mathrm{Var}(Y) + 2\alpha \cdot \beta \cdot \mathrm{cov}(X, Y).$$

Notons que si la variance d'une variable aléatoire est une constante toujours positive, la covariance de deux variables peut être négative. Si les variables X et Y sont non corrélées, c'est-à-dire que $\mathrm{cov}(X, Y) = 0$, ce qui se produit en particulier si elles sont indépendantes, on obtient :

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Si $X_1, \cdots, X_n$ sont n copies indépendantes d'une même variable aléatoire X, on a :

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X) \quad \text{et} \quad \mathrm{Var}(X_1 + \cdots + X_n) = n \cdot \mathrm{Var}(X).$$

En conséquence, la variable aléatoire que constitue la moyenne empirique vérifie :

$$\mathbb{E}(\bar{X}) := \mathbb{E}\left[\frac{1}{n}(X_1 + \cdots + X_n)\right] = \mathbb{E}(X) \quad \text{et} \quad \mathrm{Var}(\bar{X}) = \frac{1}{n}\mathrm{Var}(X).$$

2 Lois de probabilités de variables aléatoires

2.1 Loi de probabilité et fonction de répartition

Par la suite, nous dirons qu'une variable aléatoire X suit la *loi de probabilité* $\mathbb{P}_X$ (ce que l'on notera également $X \sim \mathbb{P}_X$) lorsque pour tout ensemble A borélien de $\mathbb{R}$:

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A).$$

On dira que cette loi admet une *densité* f_X par rapport à la mesure de Lebesgue lorsque l'on pourra écrire tout ensemble A borélien de $\mathbb{R}$:

$$\mathbb{P}_X(A) = \int_A f_X(x)\,dx.$$

On peut également définir la *fonction de répartition* F_X de la variable X, telle que $\forall x \in \mathbb{R}$,

$$F_x(x) = \mathbb{P}(X \leq x) = \mathbb{P}_X(]-\infty, x]).$$

Cette fonction $F_X : \mathbb{R} \to [0,1]$ est une fonction croissante, continue à droite et ayant une limite à gauche, telle que :

$$\lim_{x\to-\infty} F_X(x) = 0 \quad \text{et} \quad \lim_{x\to+\infty} F_X(x) = 1.$$

Il y a une correspondance bijective entre la connaissance de $\mathbb{P}_X$ et celle de F_X. La fonction de répartition permet également de définir les quantiles qui sont essentiels à la construction d'intervalles de confiance et de test.

2.2 Quantiles d'une loi de probabilité

Soit $\alpha \in [0,1]$. Des propriétés de la fonction de répartition, on en déduit qu'il existe $x_\alpha \in \mathbb{R}$, tel que :

$$\lim_{x\to x_\alpha} F_X(x) \leq \alpha \leq F_X(x_\alpha). \tag{A.1}$$

Soit $I_\alpha = \{x_\alpha \in \mathbb{R} \text{ tel que } x_\alpha \text{ vérifie } (A.1)\}$. On appelle *quantile* (ou fractile, ou percentile en anglais) d'ordre α de la loi $\mathbb{P}_X$, noté q_α, le milieu de l'intervalle I_α. Evidemment, lorsque X admet une distribution absolument continue par rapport à la mesure de Lebesgue, $q_\alpha = F_X^{-1}(\alpha)$, où F_X^{-1} désigne la fonction réciproque de F_X.

Deux cas particuliers sont à connaître :

1/ pour $\alpha = 0.5$, $q_{0.5}$ est appelé la *médiane* de $\mathbb{P}_X$;

2/ pour $\alpha = 0.25$ et $\alpha = 0.75$ (respectivement), $q_{0.25}$ et $q_{0.25}$ sont appelés premier et troisième *quartile* (respectivement) de $\mathbb{P}_X$.

2.3 Principales lois utilisées en modèle linéaire

Dans le cadre du modèle linéaire, nous allons essentiellement utiliser des lois de probabilités possédant une densité par rapport à la mesure de Lebesgue :

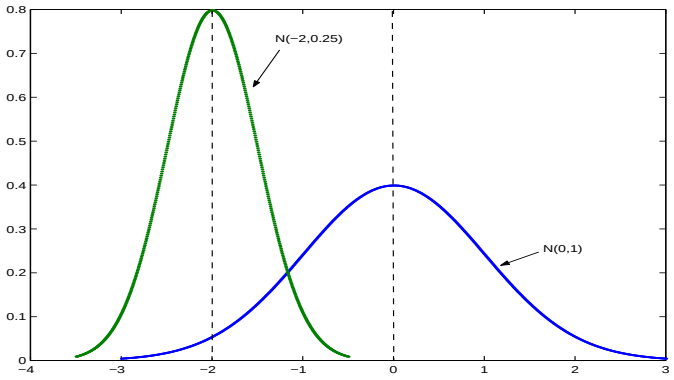


FIGURE A.1 – Représentation de la densité de la loi $\mathcal{N}(0,1)$ et de la loi $\mathcal{N}(-2, 0.5^2)$

Loi normale (ou gaussienne) centrée réduite :

Cette loi est généralement notée $\mathcal{N}(0,1)$. C'est la loi de probabilité à valeurs dans $\mathbb{R}$ de densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

On a :

$$\mathbb{E}(X) = 0 \quad \text{et} \quad \text{Var}(X) = 1.$$

Loi normale (ou gaussienne) de moyenne m et de variance σ^2 :

Si Z suit la loi $\mathcal{N}(0,1)$, $X = m + \sigma Z$ suit par définition la loi $\mathcal{N}(m, \sigma^2)$, loi normale d'espérance m et de variance σ^2. La densité de X est donnée par :

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

La figure A.1. représente la densité de la loi normale centrée réduite et celle d'une loi normale non centrée et non réduite.

A partir de la loi gaussienne, on peut en déduire les lois suivantes.

Loi du χ^2 à n degrés de libertés :

Soit $X_1, \cdots, X_n$, n variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$, alors

$$S = X_1^2 + \cdots + X_n^2$$

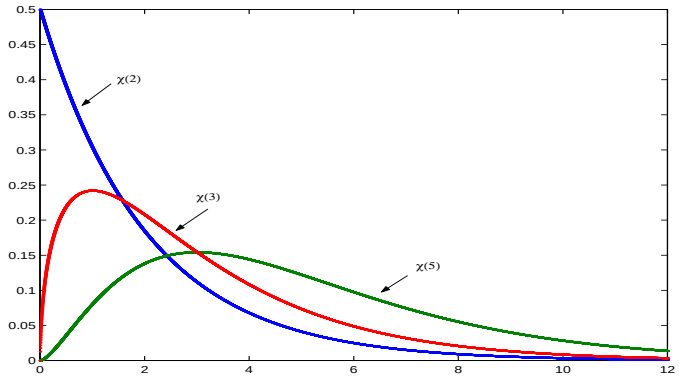


FIGURE A.2 – Représentation de la densité des lois $\chi^2(2)$, $\chi^2(3)$ et $\chi^2(5)$

suit une loi du χ^2 à n degrés de libertés, loi notée $\chi^2(n)$. Cette loi est à valeurs dans $\mathbb{R}_+$, d'espérance n et de variance $2n$. C'est aussi la loi Gamma de paramètres $(n/2, 1/2)$, c'est-à-dire que $X \sim \chi^2(n)$ admet pour densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{2^{n/2} \cdot \Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{x}{2}\right) \cdot \mathbb{1}_{\{x \geq 0\}},$$

où la fonction Gamma est telle que $\Gamma(a) = \displaystyle\int_0^\infty x^{a-1} \cdot e^{-x}$ pour $a \geq 0$. Enfin, si X suit une loi $\chi^2(n)$, par définition on dira que $Y = \sigma^2 \cdot X$ suit une loi $\sigma^2 \cdot \chi^2(n)$. La figure A.2. exhibe trois tracés différents de densité de loi du χ^2.

Loi de Student à n degrés de libertés :

La loi de Student à n degrés de liberté, notée $T(n)$, est la loi du quotient

$$T = \frac{N}{\sqrt{S/n}}$$

où N suit une loi $\mathcal{N}(0,1)$ et S suit une loi $\chi^2(n)$, N et S étant deux variables aléatoires indépendantes. Il est également possible de déterminer la densité d'une telle loi par rapport à la mesure de Lebesgue, à savoir,

$$f_X(x) = \frac{1}{\sqrt{n} \cdot B(1/2, n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2},$$

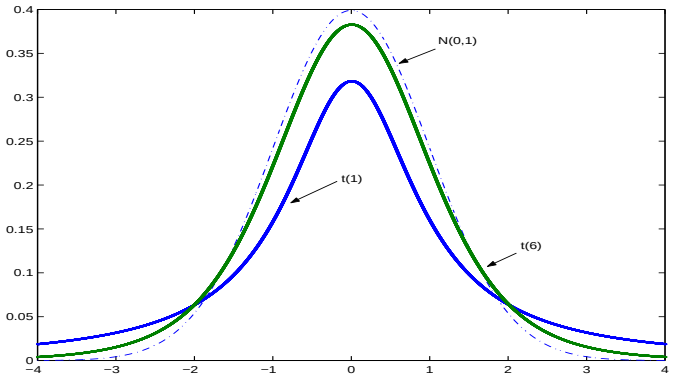


FIGURE A.3 – Représentation de la densité des lois $T(2)$, $T(6)$, à comparer avec celle de la loi $\mathcal{N}(0,1)$

où la fonction Beta est telle que $B(a,b) = \dfrac{\Gamma(a)\cdot\Gamma(b)}{\Gamma(a+b)}$ pour $a>0$ et $b>0$. La figure A.3. illustre deux exemples de cette densité, que l'on compare également avec la densité de la loi normale centrée réduite.

Remarque : Par la loi des grands nombres, plus n est grand, plus S est proche de son espérance qui vaut n. Le dénominateur est donc proche de 1. Il s'ensuit que la loi $T(n)$ est d'autant plus proche d'une loi normale que n est grand.

Un des principaux intérêt de la loi de Student réside dans le fait que si $X_1,\cdots,X_n$ sont n variables aléatoires indépendantes de loi $\mathcal{N}(m,\sigma^2)$, si on considère la moyenne et la variance empiriques :

$$\bar{X}_n = \frac{1}{n}(X_1+\cdots+X_n) \quad \text{et} \quad \bar{\sigma}_n^2 = \frac{1}{n-1}\left((X_1-\bar{X}_n)^2+\cdots+(X_d-\bar{X}_n)^2)\right),$$

alors

$$T = \frac{\sqrt{n}\cdot(\bar{X}_n - m)}{\sqrt{\bar{\sigma}_n^2}}$$

suit une loi de Student à $(n-1)$ degrés de liberté.

Loi de Fisher à n_1 et n_2 degrés de liberté :

Soit S_1 et S_2 deux variables aléatoires indépendantes de loi respectives $\chi^2(n_1)$ et $\chi^2(n_2)$. Alors par définition :

$$F = \frac{S_1/n_1}{S_2/n_2}$$

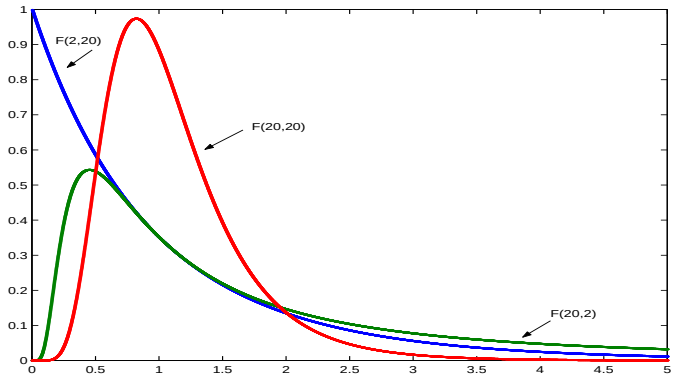


FIGURE A.4 – Représentation de la densité des lois $F(2,20)$, $F(20,2)$ et $F(20,20)$

suit une loi de Fisher à n_1 et n_2 degrés de liberté, notée $F(n_1, n_2)$.

Remarque : Par les mêmes considérations que précédemment, la loi F est d'autant plus proche de 1 que les degrés de liberté n_1 et n_2 sont grands.

On a également les propriétés suivantes :

- Si F suit une loi $F(n_1, n_2)$, alors la loi de $\dfrac{n_1}{n_2} F$ est une loi beta de seconde espèce de paramètres $(n_1/2, n_2/2)$, c'est-à-dire que F est à valeurs dans $\mathbb{R}_+$ et admet la densité par rapport à la mesure de Lebesgue :

$$f_X(x) = \frac{1}{B(n_1/2, n_2/2)} n_1^{n_1/2} \cdot n_2^{n_2/2} \frac{x^{n_1/2-1}}{(n_2 + n_1 \cdot x)^{(n_1+n_2)/2}} \mathbb{1}_{\{x \geq 0\}},$$

  la notation B désignant encore la fonction Beta.
- Si $F \sim F(n_1, n_2)$, alors $\mathbb{E}(F) = \dfrac{n_2}{n_2 - 2}$ lorsque $n_2 > 2$ et $\mathrm{Var}(F) = \dfrac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2}$ lorsque $n_2 > 4$.
- Si T suit une loi de Student $T(n)$, alors T^2 suit une loi de Fisher $F(1, n)$.

La figure A.4. donne une idée de la distribution d'une loi de Fisher pour différents choix des paramètres.

3 Vecteurs aléatoires

On se place en général dans la base canonique orthonormale de $\mathbb{R}^d$. Si Z est un vecteur aléatoire à valeurs sur $\mathbb{R}^d$, on définit $\mathbb{E}(Z)$, le vecteur dont les coordonnées sont les espérances des coordonnées de Z. Ainsi, si dans la base canonique de $\mathbb{R}^d$, $Z = (Z_1, \cdots, Z_d)'$,

$$\mathbb{E}(Z) = \mathbb{E}\begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}(Z_1) \\ \vdots \\ \mathbb{E}(Z_d) \end{pmatrix}.$$

De la même manière, on définira l'espérance d'une matrice dont les coordonnées sont des variables aléatoires par la matrice dont les coordonnées sont les espérances de chacune de ces variables aléatoires.

Ceci nous permet de définir la matrice de variance-covariance de Z de la manière suivante :

$$\mathrm{Var}(Z) = \mathbb{E}\left[(Z - \mathbb{E}(Z)).(Z - \mathbb{E}(Z))'\right]$$

donc si $Z = (Z_1, \cdots, Z_d)'$,

$$\mathrm{Var}\begin{pmatrix} Z_1 \\ \vdots \\ Z_d \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdots & \mathrm{Cov}(Z_1, Z_d) \\ \mathrm{Cov}(Z_1, Z_2) & \mathrm{Var}(Z_2) & \cdots & \mathrm{Cov}(Z_2, Z_d) \\ : & : & \cdots & : \\ \mathrm{Cov}(Z_1, Z_d) & \mathrm{Cov}(Z_2, Z_d) & \cdots & \mathrm{Var}(Z_d) \end{pmatrix}$$

matrice (d, d) dont les éléments diagonaux sont les variances et les éléments non diagonaux sont les covariances des coordonnées de Z (remarquons que la variance de Z_1 est aussi la covariance de Z_1 et de Z_1).

On vérifie également le résultat suivant : si C est une matrice (p, d) à coordonnées constituées de réels constants et si Z est un vecteur aléatoire à valeurs dans $\mathbb{R}^d$, alors $C \cdot Z$ est un vecteur de taille p de matrice de variance-covariance

$$\mathrm{Var}(C \cdot Z) = C \cdot \mathrm{Var}(Z) \cdot C'.$$

En particulier, si p vaut 1, alors $C = h'$ où h est un vecteur de taille d, et :

$$\mathrm{Var}(h' \cdot Z) = h' \cdot \mathrm{Var}(Z) \cdot h.$$

Notez que cette dernière quantité est un scalaire.

4 Vecteurs gaussiens

Soit $Y_1, \cdots, Y_d$ des variables aléatoires indépendantes de même loi $\mathcal{N}(0, \sigma^2)$, indépendantes (ce qui, dans le cas gaussien, est équivalent à $\text{cov}\,(Y_i, Y_j) = 0$ pour $i \neq j$). On considère le vecteur $Y = (Y_1, \cdots, Y_d)'$. En raison de l'indépendance, Y est un vecteur gaussien admettant une densité f_Y (par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$) qui est le produit des densités de chacune des coordonnées, soit :

$$\begin{aligned} f_Y(y_1, \cdots, y_d) &= f_{Y_1}(y_1) \times f_{Y_2}(y_2) \times \cdots \times f_{Y_d}(y_d) \\ &= \left(2\pi\sigma^2\right)^{-d/2} \exp\left(-\frac{1}{2\sigma^2}(y_1^2 + \cdots + y_d^2)\right) \\ &= \left(2\pi\sigma^2\right)^{-d/2} \exp\left(-\frac{\|y\|^2}{2\sigma^2}\right), \end{aligned}$$

avec $y = (y_1, \cdots, y_d)$. On voit donc que la densité de Y ne dépend que de la norme $\|Y\|$: elle est constante sur toutes les sphères centrées en zéro. Cela implique qu'elle est invariante par rotation ou symétrie orthogonale d'axe passant par 0 : elle est invariante par toutes les isométries de $\mathbb{R}^d$: on dira que Y suit une loi gaussienne isotrope. Rappelons que les isométries correspondent à des changements de bases orthonormées (BON). En conséquence, on a la première propriété importante :

Propriété A.1 *Soit Y un vecteur aléatoire de $\mathbb{R}^d$ de loi normale isotrope variance σ^2, c'est-à-dire que dans une BON les coordonnées de Y vérifient $\mathbb{E}(Y) = 0$ et $Var(Y) = \sigma^2 \cdot Id$. Alors les coordonnées de Y dans toute BON sont encore des lois $\mathcal{N}(0, \sigma^2)$ indépendantes.*

Voici maintenant l'un des résultats (encore appelé Théorème de Cochran) que nous utilisons le plus et nous en donnons donc une démonstration.

Théorème A.1 (Théorème de Cochran) *Soit E_1 et E_2, deux sous-espaces vectoriels orthogonaux de $E = \mathbb{R}^d$ de dimensions respectives k_1 et k_2 et soit Y un vecteur aléatoire de $\mathbb{R}^d$ de loi normale centrée isotrope de variance σ^2. Alors $P_{E_1}(Y)$ et $P_{E_2}(Y)$ sont deux variables aléatoires gaussienne centrées indépendantes et $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2 \cdot \chi^2(k_1)$ (resp. $\sigma^2 \cdot \chi^2(k_2)$). Ce théorème se généralise naturellement pour $2 < m \leq d$ sous-espaces vectoriels orthogonaux $(E_i)_{1 \leq i \leq m}$ de $E = \mathbb{R}^d$.*

Démonstration : Soit $(e_1, \cdots, e_{k_1})$ et $(e_{k_1+1}, \cdots, e_{k_1+k_2})$ deux BON de E_1 et E_2 (respectivement). L'ensemble de ces deux bases peut être complété en

$$(e_1, \cdots, e_{k_1}, e_{k_1+1}, \cdots, e_{k_1+k_2}, e_{k_1+k_2+1}, \cdots, e_d)$$

pour former une BON de $\mathbb{R}^d$ (du fait que E_1 et E_2 sont orthogonaux).

Soit $(Y_1, \cdots, Y_d)$, les coordonnées de Y dans cette base ; elles sont indépendantes de loi $\mathcal{N}(0, \sigma^2)$ car le changement de base est orthonormal et nous avons vu que la distribution de Y était conservé par transformation isométrique. Comme

$$P_{E_1}(Y) = Y_1 e_1 + \cdots + Y_{k_1} e_{k_1} \Longrightarrow \|P_{E_1}(Y)\|^2 = \sigma^2 \left(\left(\frac{Y_1}{\sigma} \right)^2 + \cdots + \left(\frac{Y_{k_1}}{\sigma} \right)^2 \right)$$

$$P_{E_2}(Y) = Y_{k_1+1} e_{k_1+1} + \cdots + Y_{k_1+k_2} e_{k_1+k_2} \Longrightarrow \|P_{E_2}(Y)\|^2 = \sigma^2 \left(\left(\frac{Y_{k_1+1}}{\sigma} \right)^2 + \cdots + \left(\frac{Y_{k_1+k_2}}{\sigma} \right)^2 \right).$$

On voit bien ainsi l'indépendance entre les deux projections et le fait que la loi de $\|P_{E_1}(Y)\|^2$ (resp. $\|P_{E_2}(Y)\|^2$) est une loi $\sigma^2 \cdot \chi^2(k_1)$ (resp. $\sigma^2 \cdot \chi^2(k_2)$). ■

On peut définir plus généralement un vecteur gaussien Y à valeurs dans $\mathbb{R}^d$ (non dégénéré), d'espérance $\mu \in \mathbb{R}^d$ et de matrice de variance-covariance Σ quelconques (du moment que Σ soit une matrice de Toeplitz définie positive). Cela équivaut à définir un vecteur aléatoire de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$,

$$f_Y(y) = \frac{(2\pi)^{-n/2}}{|\Sigma|} \exp\left(-\frac{1}{2} (y-\mu)'.\Sigma^{-1}.(y-\mu) \right),$$

pour $y \in \mathbb{R}^d$, et avec $|\Sigma|$ le déterminant de la matrice Σ. Remarquons une nouvelle fois que l'espérance et la variance définissent complètement la loi de probabilité d'un vecteur gaussien.

A partir des propriétés générales sur les vecteurs aléatoires, on obtient le fait que :

Propriété A.2 *Soit Y un vecteur gaussien à valeurs dans $\mathbb{R}^d$ (non dégénéré), d'espérance $\mu \in \mathbb{R}^d$ et de matrice de variance-covariance Σ. Soit C une matrice réelle de taille (p, d) où $p \in \mathbb{N}^*$. Alors $C \cdot Y$ est un vecteur gaussien tel que :*

$$C \cdot Y \sim \mathcal{N}(C \cdot \mu \,,\, C \cdot \Sigma \cdot C')$$

On en déduit les conséquences suivantes :

- si Y est un vecteur gaussien isotrope de $\mathbb{R}^d$ de variance σ^2 et h un vecteur de $\mathbb{R}^d$, alors $h' \cdot Y$ est une combinaison linéaire des coordonnées de Y tel que :

$$h' \cdot Y \text{ suit la loi } \mathcal{N}(0, \sigma^2 \cdot h' \cdot h) = \mathcal{N}(0, \sigma^2 \cdot \|h\|^2)$$

- si Y est un vecteur gaussien d'espérance μ et de matrice de variance-covariance Σ et si h un vecteur de $\mathbb{R}^d$, alors $h' \cdot Y$ est une combinaison linéaire des coordonnées de Y et :

$$h' \cdot Y \text{ suit la loi unidimensionnelle } \mathcal{N}(h' \cdot \mu, h' \cdot \Sigma \cdot h)$$

(Pour une présentation plus détaillée des notions sur les vecteurs gaussiens on peut consulter le livre P. Toulouse (1999) [59], chap.2)

5 Théorèmes limites

Quelle est la fréquence des piles lorsque l'on jette un grand nombre de fois une pièce de monnaie ? Comment un sondage peut-il permettre d'estimer le score d'un candidat à une élection ? C'est à ce genre de question que répondent les deux théorèmes limite suivants. Le premier, appelé Loi, car on a longtemps cru qu'il ressortait des lois de la nature (du même ordre que la gravitation), dit en substance que "la moyenne empirique se rapproche de plus en plus de la moyenne théorique, ou espérance, lorsque le nombre de données croît". Plus précisément, son énoncé est le suivant :

Théorème A.2 (Loi Forte des Grands Nombres)
Soit $(Z_1, \cdots, Z_n)$, n variables aléatoires indépendantes distribuées suivant une même loi d'espérance μ (donc $\mathbb{E}|Z| < +\infty$). Soit $\overline{Z}_n = \frac{1}{n}(Z_1 + \cdots + Z_n)$ la moyenne empirique. Alors quand n est grand

$$\overline{Z}_n \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} \mu.$$

Notons que les hypothèses peuvent être étendues à des variables qui ne sont plus indépendantes entre elles, ou qui n'ont pas forcément la même espérance.

Le second théorème précise en quelque sorte la manière dont la moyenne empirique se rapproche de l'espérance :

Théorème A.3 (Théorème de la Limite Centrale)
Soit $(Z_1, \cdots, Z_n)$, n variables aléatoires indépendantes distribuées suivant une même loi d'espérance μ et de variance σ^2. Soit $\overline{Z}_n = \frac{1}{n}(Z_1 + \cdots + Z_n)$ la moyenne empirique. Alors quand n est grand

$$\sqrt{n}\left(\frac{\overline{Z}_n - \mu}{\sigma}\right) = \frac{\overline{Z}_n - \mu}{\sqrt{\sigma^2/n}} \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,1).$$

La traduction "intuitive" de ce résultat est que "la moyenne empirique tend à être gaussienne" quand le nombre de données devient grand. Concrètement, cela veut dire que lorsque n devient grand, $\overline{Z}_n$ a une loi très proche de la loi gaussienne de moyenne μ et de variance σ^2/n. Une application de ce résultat est par exemple la simulation informatique de variables aléatoires gaussiennes que l'on obtient en simulant la moyenne de plusieurs dizaines de variables uniformes sur $[0,1]$ (que les calculatrices ou ordinateurs savent tous (à peu près) simuler : c'est notamment la touche RAND).

Bibliographie

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[2] H. Akaike. A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.*, 30(1) :9–14, 1978.

[3] T. Amemiya. *Advanced Econometrics*. Harvard University Press, Cambridge, 1985.

[4] J.-M. Azaïs, R. A. Bailey, and H. Monod. A catalogue of efficient neighbour-designs with border plots. *Biometrics*, Biometrics(49) :1252–1261, 1993.

[5] J.-M. Azaïs. Analyse de la variance non orthogonale, l'exemple de sas/glm. *Rev. Stat. Appli.*, XLII :27–41, 1994.

[6] R. Bailey. Designs on association schemes. In *Statistics and science : a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes Monogr. Ser.*, pages 79–102. Inst. Math. Statist., Beachwood, OH, 2003.

[7] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493, 2000.

[8] P. Barbe and M. Ledoux. *Probabilités*. Belin, Paris, 1999. Espace 34.

[9] A. Bardin and J.-M. Azaïs. Une hypothèse minimale pour une théorie des plans d'expériences randomisés. *Rev. Statist. Appl.*, 38(2) :21–41, 1990.

[10] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413, 1999.

[11] P. Besse. *DATA Mining I. Exploration Statistique*. `http://www.lsp.math-ups.fr/Besse/pub/Explo_stat.pdf`. Document pédagogique du L.S.P., Université Toulouse III.

[12] P. Besse. *DATA Mining II. Modélisation Statistique & Apprentissage*. `http://www.lsp.math-ups.fr/Besse/pub/Appren_stat.pdf`. Document pédagogique du L.S.P., Université Toulouse III.

[13] P. Besse. *Pratique de la modélisation Statistique.* `http://www.lsp.math-ups.fr/Besse/pub/modlin.pdf`. Document pédagogique du L.S.P., Université Toulouse III.

[14] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268, 2001.

[15] P. Bonnet and G. Lansiaux. Mémoire de maitrise. Mémoire de maitrise, Université Paul Sabatier, 1992.

[16] K. Burnham and D. Anderson. *Model selection and multimodel inference.* Springer-Verlag, New York, second edition, 2002. A practical information-theoretic approach.

[17] P. Calas, T. Rochd, P. Druilhet, and J.-M. Azaïs. In vitro adhesion of two strains of prevotella nigrescens to the dentin of the root cana :the part played by different irrigations solutions. *Journal of Endodontics*, 24(2) :112–115, 1998.

[18] W. Cochran and G. Cox. *Experimental designs.* Wiley Classics Library. John Wiley & Sons Inc., New York, second edition, 1992. A Wiley-Interscience Publication.

[19] J. Coursol. *Techniques statistiques des modèles linéaires.* cimpa, Nice, 1981.

[20] D. Dacunha-Castelle and M. Duflo. *Probabilités et statistiques. Tome 1.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1982. Problèmes à temps fixe. [Problems with fixed time].

[21] D. Dacunha-Castelle and M. Duflo. *Probabilités et statistiques. Tome 2.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Problèmes à temps mobile. [Movable-time problems].

[22] J.-B. Denis. Two-way analysis using covariates. *Statistics*, 19(1) :123–132, 1988.

[23] H. Dette and W. Studden. *The theory of canonical moments with applications in statistics, probability, and analysis.* Wiley Series in Probability and Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1997. A Wiley-Interscience Publication.

[24] N. Draper and H. Smith. *Applied regression analysis.* Wiley Series in Probability and Statistics : Texts and References Section. John Wiley & Sons Inc., New York, third edition, 1998. With 1 IBM-PC floppy disk (3.5 inch ; DD), A Wiley-Interscience Publication.

[25] P. Druilhet. *Optimalité des pans d'expériences équilibrés pour les voisinages.* PhD thesis, Université Paul Sabatier, 118 route de Narbonne F31062 Toulouse, 1995.

[26] R. Feller. *Theory of probability.* John Wiley & Sons Inc., New York, 1971. Wiley Series in Probability and Mathematical Statistics.

[27] R. Fisher. *Statistical methods for research workers*. Oliver and Boyd, London, 1925.

[28] R. Fisher. *Statistical methods for research workers*. Hafner Publishing Co., New York, 1973. Fourteenth edition—revised and enlarged.

[29] G. Furnival and W. Wilson. Regression by leaps and bounds. *Technometrics*, 16 :499–511, 1974.

[30] W. Green. *Econometrics analysis*. Prentice Hall International Editions, fourth edition, 2000.

[31] X. Guyon. *Modèle linéaire et économétrie*. Ellipse, Paris, 2001.

[32] E. Hannan. *Multiple time series*. John Wiley and Sons, Inc., New York-London-Sydney, 1970.

[33] P. Huber. *Robust statistics*. John Wiley & Sons Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.

[34] C. Hurvich and C. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2) :297–307, 1989.

[35] J. Jobson. *Applied Multivariate Data Analysis*. Springer-Verlag, New York, 1991. Springer Series in Statistics.

[36] J. Kiefer. Construction and optimality of generalized Youden designs. In *A survey of statistical design and linear models (Proc. Internat. Sympos., Colorado State Univ., Ft. Collins, Colo., 1973)*, pages 333–353. North-Holland, Amsterdam, 1975.

[37] J. Kiefer and H. Wynn. Optimum balanced block and Latin square designs for correlated observations. *Ann. Statist.*, 9(4) :737–757, 1981.

[38] W. Krzanowski. *Principles of multivariate analysis*, volume 3 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1990. A user's perspective, Corrected reprint of the 1988 edition, Oxford Science Publications.

[39] J.-M. Lecoutre and P. Tassi. *Statistique non paramétrique et robustesse*. Economica, Paris, 1987.

[40] H. Levene. Robust tests for equality of variances. In *Contributions to probability and statistics*, pages 278–292. Stanford Univ. Press, Stanford, Calif., 1960.

[41] H. Linhart and W. Zucchini. *Model selection*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1986.

[42] C. Mallows. Some comments on C_p. *Technometrics*, 15 :661–675, 1973.

[43] C. Mallows. More comments on C_p. *Technometrics*, 37(4) :362–372, 1995.

[44] P. McCullagh and J. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1983.

[45] X. Milhaud. *Statistiques*. Belin, Paris, 2001. Espace 34.

[46] J. Miller and G. Rupert. *Simultaneous statistical inference*. Springer-Verlag, New York, second edition, 1981. Springer Series in Statistics.

[47] G. Milliken and D. Johnson. *Analysis of Messy Data vol. 1 : Designed Experiments*. Van Nostrand Reinhold, New-York, 1984.

[48] G. Morrot and F. Brochet. Cognition et vin. Technical report, Agro-Montpellier, France., 1992.

[49] D. Raghavarao. *Constructions and combinatorial problems in design of experiments*. John Wiley & Sons Inc., New York, 1971. Wiley Series in Probability and Mathematical Statistics.

[50] C. Rao and Y. Wu. A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2) :369–374, 1989.

[51] S. Rey and J. Trappier. Étude du goût : comparaison de coca-cola et pepsi-cola. Mémoire de dess, Université Paul Sabatier, Toulouse, France., 2004.

[52] G. Saporta. *Probabilités, analyse des données et statistique*. Technip, Paris, 1990.

[53] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[54] S. Searle. *Linear models for unbalanced data*. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1987.

[55] G. Snedecor and W. Cochran. *Statstical methods*. The Iowa State University Press, Ames, Iowa , USA, 6th edition, 1967.

[56] M. Spivak. *A comprehensive introduction to differential geometry. Vol. V*. Publish or Perish Inc., Wilmington, Del., second edition, 1979.

[57] J. Tanner. *The Physique of the Olympic Athlete*. George Allen and Unwin, London, 1964.

[58] R. Tomassone, S. Audrain, E. Lesquoy, and C. Miller. *La régression, nouveaux regards sur une ancienne méthode statistique*. Masson, Paris, 1992.

[59] P. S. Toulouse. *Thèmes de probabilités et statistiques*. Dunod, Paris, 1999.

[60] A. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.

Table des matières

Index

Y a-t-il une différence de goût entre diverses boissons au cola ? Comment faire un béton sans bulles ? Comment évolue le prix d'une voiture d'occasion en fonction de son modèle et de son âge ? Ou encore, dans un registre plus formalisé, pourquoi existe-t-il plusieurs définitions de sommes de carrés en analyse de la variance ?

En réponse à ce type de questions, nous développons des variations sur le thème du modèle linéaire (régression linéaire, analyse de la variance, modèles mixtes et plans d'expériences). Chaque notion est introduite et illustrée à l'aide de très nombreux exemples utilisant des données réelles. Les programmes permettant leurs traitements avec les logiciels R, SAS et Splus sont présentés, tout comme leurs résultats numériques et graphiques. Pour autant nous n'esquivons pas certains développements théoriques qui permettent de donner des réponses satisfaisantes à des questions plus délicates.

Cet ouvrage est essentiellement destiné aux étudiants en master de mathématiques appliquées, de biostatistique, d'économétrie, de chimie... ou aux élèves en école de commerce ou d'agronomie. Ils pourront s'initier aux diverses notions en se référant aux nombreux exercices, exemples et applications. Les enseignants trouveront également dans ce livre une source de réflexion et un support théorique pour certaines parties plus formalisées. Enfin, les professionnels utilisant la modélisation statistique dans leur travail pourront consulter dans la plupart des chapitres les nombreux passages dévolus à la pratique du modèle linéaire, avec tous les détails de mise en œuvre informatique et un vocabulaire mathématique restreint.